TEIJA WAARAMAA-MÄKI-KULMALA

Emotions in Voice

Acoustic and perceptual analysis of voice quality
in the vocal expression of emotions

■

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Humanities of the University of Tampere,
for public discussion in the Auditorium A1
of the Main Building, Kalevantie 4, Tampere,
on May 8th, 2009, at 12 o'clock.

UNIVERSITY OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere
Department of Speech Communication and Voice Research
Finland

Distribution
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 3 3551 6055
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Juha Siro

Acta Universitatis Tamperensis 1399
ISBN 978-951-44-7666-2 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 827
ISBN 978-951-44-7667-9 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2009

To My Family

"Esse est percipi. The being of things is their perception. We know a thing when we understand it, and we understand it when we can interpret or tell what it signifies."
George Berkeley (1685-1753)

Contents

List of the original articles

Article I

Waaramaa T, Laukkanen A-M, Airas M, Alku P.

Perception of Emotional Valences and Activity Levels from Vowel Segments of Continuous Speech. Accepted for publication in Journal of Voice (4/2008). Available online 25th December 2008. Copyright Elsevier.

Article II

Toivanen J, Waaramaa T, Alku P, Laukkanen A-M, Seppänen T, Väyrynen E, Airas M.

Emotions in /a:/: A Perceptual and Acoustic study.

Logopedics Phoniatrics Vocology 2006; 31: 1, 43-48.

http://www.informaworld.com/LPV

Article III

Waaramaa T, Alku, P, Laukkanen A-M.

The role of F3 in the vocal expression of emotions.

Logopedics Phoniatrics Vocology 2006; 31: 4, 153-156.

http://www.informaworld.com/LPV

Article IV

Waaramaa T, Laukkanen A-M, Alku P, Väyrynen E.

Mono-pitched expression of emotions in different vowels.

Folia Phoniatrica et Logopaedica 2008; 60: 5, 249-255.

S. Karger AB, Basel.

Author's contribution

Article I

Perception of Emotional Valences and Activity Levels from Vowel Segments of Continuous Speech.

The person in charge of the recordings of the material for the first study was Professor Paavo Alku. The author of this dissertation participated in the recordings, edited the material, performed the acoustic analyses, edited the tape for the listening test, organized and performed the listening tests, analyzed the results of them, participated in the statistical analyses performed and wrote the manuscript. Hanna-Mari Puuska M.A. was the statistical specialist. The author created the figures together with Professor Anne-Maria Laukkanen and Professor Paavo Alku. Professor Paavo Alku also performed the inverse filtering and Matti Airas D.Sc.(Tech.) calculated the NAQ parameter. Professor Anne-Maria Laukkanen and Professor Paavo Alku provided their comments to the manuscript.

Article II

Emotions in /a:/: A Perceptual and Acoustic study.

The material for the second article was derived from the data of the first study, and thus it was analyzed for the most part both acoustically and perceptually by the author of this dissertation. The automatic classification experiments were performed by the MediaTeam in the University of Oulu by Professor Tapio Seppänen, Juhani Toivanen Ph.D. and Eero Väyrynen M.Sc. The results of the listening test originally performed for the first study were here compared with the results of the automatic classification. The present author wrote the chapters of the manuscript concerning the acoustic analyses, the human listening test and its results, the first author being Juhani Toivanen Ph.D. Professor Anne-Maria Laukkanen and Professor Paavo Alku contributed the manuscript with

their comments.

Article III

The role of F3 in the vocal expression of emotions.

The material for the third study was also derived from the first investigation. The present author chose the best recognized samples in the human listening test performed for the first study. Professor Paavo Alku modified the samples, and the present author edited the tape for the listening test, organized and performed the listening tests for the semi-synthesized samples, analyzed the results of the listening tests, performed the statistical analyses, created the figures together with Professor Paavo Alku and wrote the article. Professor Paavo Alku and Professor Anne-Maria Laukkanen provided their comments on the manuscript.

Article IV

Mono-pitched expression of emotions in different vowels.

The present author analyzed the vowel samples recorded for the fourth study for the main acoustic parameters used. Professor Paavo Alku inverse filtered the samples and calculated the NAQ parameter. The listening tape was edited and the listening tests were organized and performed by the author. The statistical analyses were performed by Hanna-Mari Puuska M.A. and partly by the author assisted by Professor Anne-Maria Laukkanen and Eero Väyrynen M.Sc. The figures were created together with Professor Paavo Alku, Professor Anne-Maria Laukkanen and the present author. The manuscript was written by the present author with the support of the comments from Professor Anne-Maria Laukkanen and Professor Paavo Alku.

Teija Waaramaa-Mäki-Kulmala	Anne-Maria Laukkanen
The Author	The Principal Supervisor

List of abbreviations and symbols

ANCOVA	Analysis of Covariance
ClQ	Closing quotient
CQ	Closed quotient
dB	Decibell
d_{peak}	Negative peak of the first derivative of the flow waveform
EGG	Electroglottography
F0	Fundamental frequency
F1	First formant
F2	Second formant
F3	Third formant
F4	Fourth formant
f_{ac}	Altering current flow
FFT	Fast-Fourier Transform, normalized average spectrum
Hz	Hertz
IAIF	Iterative Adaptive Inverse Filtering
ISA	Intelligent Speech Analyser
kHz	Kilohertz
L_{eq}	Equivalent sound level
LTAS	Long-term-average spectrum
ms	Millisecond
NAQ	Normalized amplitude quotient
OQ	Open quotient
SD	Standard deviation

S/N	Signal-to-noise ratio
SPL	Sound pressure level
SQ	Speed quotient
T	Time, period length
VAS	Visual analog scale

Finnish summary/Artikkeliväitöskirjan tiivistelmä

Teija Waaramaa-Mäki-Kulmala

Emotions in voice
Acoustic and perceptual analysis of voice quality in the vocal expression of emotions

Emootiot äänessä
Tunneilmaisun akustiset ominaisuudet ja vastaanotto

1. Johdanto

Inhimillinen kommunikaatio sisältää aina emotionaalista informaatiota halusimme sitä tai emme. Ei ainoastaan puheen sisältö vaan myös puhujan ulkoiset ja äänen ominaisuudet välittävät viestiä. Vastaanottaja kokee useimmiten neutraaliksi tarkoitetun viestin negatiiviseksi, sillä puhe ilman minkäänlaista positiivista sävyä ymmärretään helposti tylyksi. Puhujasta syntyneiden vaikutelmien perusteella kuulijat päättävät omasta asennoitumisestaan puhujaan ja kuulemaansa. Sekä puhujan että vastaanottajan asenteisiin ei kuitenkaan vaikuta pelkästään kulloinenkin tilanne, vaan myös heidän yksilölliset fyysiset ja psyykkiset ominaisuutensa sekä heidän aiemmat kokemuksensa ja käsityksensä. Pienilläkin puheen vivahde-eroilla voi olla tärkeä merkitys henkilökohtaisten assosiaatioiden syntymiselle. Näitä nyansseja on luonnollisesti sitä helpompi havaita, mitä paremmin viestijät tuntevat toisensa. Emotionaalista kokemusmaailmaa ja sen muutoksia voidaan kuvailla hermeneuttisen kehän tai spiraalin käsitteellä, jossa jo aiemmin koettuun tai opittuun lisätään uutta. Uusi muokkaa vanhaa ja vastavuoroisesti vanha vaikuttaa siihen, miten uusi koetaan. Tällaisten henkilökohtaisten ominaisuuksien sekä oman ympäröivän kulttuurin konventioiden lisäksi kannamme mukanamme myös universaaleja evoluution muokkaamia tapoja ilmaista ja vastaanottaa tunteita. Tunteiden ilmaisun ja vastaanoton eri

kerrostumien universumia voidaan kutsua ehdotukseni mukaan emosfääriksi. Emosfääri muodostaa täten nelikentän:

Emosfäärin nelikenttä

Universaali intrapersoonallinen	Universaali interpersoonallinen
Kulttuurinen intrapersoonallinen	Kulttuurinen interpersoonallinen

Tunteet voidaan jakaa perustunteisiin ja sosiaalisiin eli kulttuurisiin tunteisiin. Englanninkielisten käsitteiden *affect* ja *emotion* käytössä on ollut suurta häilyvyyttä, eikä näiden käsitteiden välillä yleensä tehdä eroa. Tässä tutkimuksessa on kuitenkin pyritty erittelemään niiden merkitykset niin, että *affect* saa merkityksen perustunne (lat. *afficere/affectum* = ärsykkeen aiheuttama mielentila) ja *emotion* (lat. *emotio*: e- = pois, *movere*/*motum* = liikkua, liikuttaa) saa merkityksen sosiaalinen/kulttuurinen tunne. Perustunteet ilo, suru, viha ja pelko sekä usein lisäksi myös yllätys ja inho ovat universaaleja ja niihin liittyy selkeä fyysinen reaktio, jota voidaan pitää ensisijaisena perustunteiden ilmenemismuotona. Universaalius merkitsee tunteen samankaltaista fyysistä kokemista kaikkialla maailmassa, vaikkakin tunteen ilmaisutapa voi saada kulttuurisia ominaispiirteitä. Esimerkiksi surua saatetaan ilmaista joko surullisella ilmeellä tai hymyillen kulttuurista riippuen. Perustunteet esiintyvät tuskin koskaan yksin ainoana tunteena, jonka ihminen kokee kulloisellakin hetkellä. Sen sijaan sosiaalisessa kontekstissa tunteilla on taipumus muodostaa keskenään erilaisia kombinaatioita, jotka vaikuttavat toisiinsa. Toiset tunteista ovat lyhytkestoisia, toiset pitempikestoisia. Näin muotoutuneita tunteita kutsutaan sosiaalisiksi tai kulttuurisiksi tunteiksi, joita voivat olla esimerkiksi empatia, rakkaus, häpeä tai epävarmuus. Sillä, minkälaiseksi kunkin yksilön emosfääri on muovautunut, on puolestaan suuri merkitys ihmisen ns. taustatunnetilaan. Taustatunne on yksilön yleistä emootionaalista asennoitumista kuvaava termi:

toiset suhtautuvat ihmisiin ja asioihin yleisesti esimerkiksi positiivisesti, toiset skeptisesti, toiset negatiivisesti. Emosfäärin kerrostuneisuuden takia käytetään tämän tutkimuksen englanninkielisessä tekstissä termiä *emotion* (ei termiä *affect*) ja vastaavasti tässä suomenkielisessä tiivistelmässä joko termiä emootio tai sen suomenkielistä vastinetta *tunne*.

Termien affekti ja emootio määritelmät ja niiden erot

Affekti	Emootio
- primaarinen (tai perus) - aivojen vanhemmat osat - ei ole opeteltavissa - biofysiologinen reaktio ärsykkeeseen - akuutti - universaali - voi johtaa emootioihin - tahdosta riippumaton	- sekundaarinen - aivojen nuoremmat osat - voidaan oppia - kognitiivinen reaktio ärsykkeeseen - lyhyt- tai pitkäkestoinen - kulttuurisidonnainen tai sosiaalinen - voi saada aikaiseksi uusia emootioita (tai niiden yhdistelmiä) - motivoitu, jokseenkin tahdonalainen

2. Tutkimuskysymykset

Tämän tutkimuksen tavoitteena oli selvittää, mitä vaikutuksia äänenlaadun eri akustisilla ominaisuuksilla on emootioiden välittämisessä. Tavoitteena oli lisäksi saada selville, voiko pelkän luennasta eristetyn pääpainollisen vokaalin mittaisesta signaalista havaita emootiota tai sen valenssia, ja voiko vokaalista, jossa sävelkorkeus pidetään samana, tunnistaa eri tunnetiloja. Tähän mennessä puheeseen liittyvän tunneilmaisun tutkimuksessa pääpaino on ollut perustaajuudessa, voimakkuudessa, niiden lauseensisäisessä vaihtelussa sekä puhe- ja artikulaatiotempossa. Tiedetään, että äänen perustaajuudella (F0, Hz) ja äänenpainetasolla (SPL tai sen aikakeskiarvo L_{eq}, dB) sekä ilmaisun kestosuhteilla on vaikutusta emootioiden välittämisessä. Sitä vastoin äänenlaadun roolia emootioiden välittämisessä ja välittymisessä on tutkittu erittäin vähän. Akustisesti äänenlaatu

ilmenee ennen kaikkea puhesignaalin energian erilaisena jakautumisena taajuusasteikolle. Tätä puolestaan voidaan tarkastella kahdella tasolla, äänilähteen (äänihuuliväräh telyn tuottama periodinen ilmavirtausvaihtelu) ja suotimen (ääniväylän resonanssit eli formantit, jotka vaihtelevat artikulaation mukaan) tasolla. Tämän tutkimuksen tuloksia voidaan soveltaa äänen ja puhetekniikan kouluttamisessa. Lisäksi tietoa emotionaalisen ilmaisun äänellisistä piirteistä voidaan hyödyntää puheteknologiassa, kuten puhujan- ja puheentunnistuksen ja synteesin kehittämisessä. Näitä voidaan parantaa, kun käytettävissä on enemmän tietoa individuaalisista variaatiomahdollisuuksista äänisignaalissa ja niiden perkeptuaalisesta merkityksestä.

3. Tutkimusaineisto ja –menetelmät

Tutkimus koostuu neljästä osatutkimuksesta. Materiaalina ensimmäisessä ja toisessa osatutkimuksessa käytettiin ammattinäyttelijöiden (N 9) tuottamia ääninäytteitä. Näyttelijät lukivat minuutin mittaisen kappaleen proosatekstiä tarkoituksenaan ilmaista iloa, surua, hellyyttä, vihaa ja neutraalia tunnetilaa. Luennasta editoitiin pitkä [a:]-vokaali sanasta *tAAkkahan* (vokaalin kesto ~ 150 ms) kuuntelukoetta varten. Kuuntelukokeeseen osallistui 50 kuuntelijaa, joiden tehtävänä oli vastata, mitä tunteita he kuulivat vokaalinäytteistä. Vastaukset luokiteltiin niiden valenssin mukaan. Näytteistä (N 171) mitattiin F0, F1, F2, F3, F4 (neljä alinta formanttitaajuutta), L_{eq}, näytteen kesto, alfa ratio ja NAQ (normalized amplitude quotient). Alfa ratio mittaa äänienergian jakautumista 1 kHz:n ylä- ja alapuolelle, ja siten se heijastaa ääntötapaa. Hyperfunktionaalisessa, puristeiselta kuulostavassa äänentuotossa spektrin kaltevuus on loiva. Hypofunktionaalisessa, pehmeältä, vuotoiselta kuulostavassa äänentuotossa spektri on puolestaan jyrkkä. Tässä tutkimuksessa alfa ratio laskettiin seuraavasti: L_{eq} (50 Hz-1 kHz) – L_{eq} (1-5 kHz). Lähdeäänen ja ääniväylän vaikutusten erottamisessa käytettiin IAIF (Iterative Adaptive Inverse Filtering) –käänteissuodatusmenetelmää,

jossa akustisesta äänenpainesignaalista estimoidaan ääniväylän resonanssien ja huulisäteilyn vaikutukset. Niiden kumoamisella saadaan estimaatti lähdeäänestä. Lähdeäänen kuvaamisessa käytettiin normalisoitua amplitudisuhdelukua (normalized amplitude quotient, NAQ), joka kertoo niin ikään ääntötavasta. NAQ lasketaan seuraavalla tavalla: NAQ = $f_{ac}/(d_{peak}T)$. Kaavassa f_{ac} on virtauspulssin amplitudi ja d_{peak} on virtaussignaalin derivaatan minimi, joka kertoo virtauksen vähenemisen nopeudesta eli heijastaa ääniraon sulkeutumisnopeutta. T on periodin kesto. Tilastoanalyysissä käytettiin Multinomial Logistic Regression Analysis –menetelmää selvittämään vastaanottoon vaikuttaneiden parametrien välisiä suhteita. Muuttujien keskinäisiä suhteita tarkasteltiin Pearsonin korrelaatiolla.

Toisessa osatutkimuksessa verrattiin ensimmäisen osatutkimuksen kuuntelukokeen tuloksia varsinaisten emootioiden tunnistamisen (ei valenssin) osalta tietokonetunnistukseen. Emootioiden tunnistusprosenttimääriä verrattiin ristiintaulukoinnin avulla. Kuuntelukokeessa merkitseviksi osoittautuneita parametreja verrattiin tietokonetunnistuksessa tilastollista merkitsevyyttä saaneisiin parametreihin. Tietokonetunnistus pohjautui kNN (k-Nearest Neighbor classifier) –menetelmään.

Kolmannessa osatutkimuksessa tarkasteltiin F3:n merkitystä emotionaalisen valenssin vastaanotossa. Materiaaliksi valittiin ensimmäisessä kuuntelukokeessa hyvin tunnistetut kolme näytettä, jotka edustivat surua, vihaa ja hellyyttä. Näytteet syntetisoitiin IAIF–analyysin pohjalta siten, että F3:n taajuutta nostettiin ja laskettiin 30 %:a suhteessa alkuperäiseen taajuuteen sekä poistettiin F3 kokonaan. Syntetisoidut näytteet (N 12) esitettiin 30 kuuntelijalle, jotka arvioivat näytteiden valenssia visuaalis-analogisella VAS–asteikolla, jonka ääripäät olivat negatiivinen (0 mm) ja positiivinen (100 mm) sekä keskikohta neutraali (50 mm). Esimerkiksi ilolla on positiivinen valenssi ja surulla negatiivinen. Silloin, kun ei tarkoituksellisesti ilmaista mitään erityistä tunnetta, katsotaan, että kyseessä on neutraali valenssi. Alkuperäisten ja modifioitujen näytteiden vastaanoton eroja tarkasteltiin Wilcoxon Signed Rank t-testillä. Vastaanotetun valenssin ja F3:n taajuuden välistä suhdetta tutkittiin Pearsonin korrelaatiolla.

Neljännessä osatutkimuksessa käytettiin materiaalina näyttelijäopiskelijoiden (N 13) tuottamia pitkiä vokaaleja [a:], [i:] ja [u:]. Jokainen puhuja tuotti vokaalit (N 195) omalta habituaaliselta puhekorkeudeltaan pitäen sävelkorkeuden samana ilmaisun keston ajan. Tuotetut tunnetilat olivat ilo, suru, viha, hellyys ja neutraali tunne. Näytteiden kesto oli ~2400 ms. Kuuntelukokeeseen osallistui 20 naista ja 20 miestä, joiden tehtävänä oli tunnistaa tuotetut emootiot. Vastauksista tutkittiin, onko sukupuolten välillä eroja emootioiden vastaanotossa. Näytteet analysoitiin mittaamalla niistä akustiset parametrit F0, F1, F2, F3, F4, L_{eq}, kesto, alfa ratio ja NAQ. NAQ mitattiin pelkästään [a:]:sta. Aineisto analysoitiin ANCOVA (SPSS15) –varianssianalyysillä. Muiden parametrien riippuvuutta L_{eq}:stä tutkittiin asettamalla L_{eq} kovariaatiksi. Riippuviksi muuttujiksi valittiin filtteri (formanttitaajuudet), NAQ, L_{eq} ja alfa ratio.

Kaikkien osatutkimusten akustiset analyysit NAQ:ia lukuun ottamatta tehtiin ISA (Intelligent Speech Analyser) – ohjelmalla, jonka on kehittänyt DI Raimo Toivonen. Osatutkimuksissa I ja III-IV emootiot luokiteltiin niiden valenssin (positiivinen, neutraali, negatiivinen) ja niiden oletetun psykofyysisen aktiviteettitason (korkea, keskiverto, matala) mukaan, minkä avulla nämä kaksi muuttujaa voitiin koodata tilastollista käsittelyä varten. Ilolle ja vihalle määriteltiin korkea aktiviteettitaso, surulle ja hellyydelle matala sekä neutraalille keskiverto aktiviteettitaso. Kaikissa kuuntelukokeissa tutkittiin intra- ja interraterreliabiliteetti joko prosentuaalisina osuuksina tai Cronbachin alfalla.

4. Tärkeimmät tutkimustulokset ja johtopäätökset

1. Lyhyet ~150 ms:n kestoiset vokaalinäytteet osoittautuivat riittävän pitkiksi välittääkseen emootioiden valenssia ja ~2400 ms:n kestoiset yhdeltä taajuudelta ilmaistut näytteet riittävän mittaisiksi emootioiden nimeämiseksi.

2. Luennasta editoitujen pääpainollisten pitkien [a:]-vokaalien automaattinen tilastopohjainen emootioluokittelu antoi paremman tuloksen kuin koehenkilöiden kuuntelukoe. Tietokone ei kuitenkaan tunnistanut yhtä hyvin vihaa kuin koehenkilöt, mutta se taas tunnisti paremmin ilon kuin kuuntelijat. Tietokonetunnistus näytti hyödyntävän pääasiassa eri parametrejä (kesto ja alfa ratio sekä S/N, NAQ ja jitterin keskiarvo) emootioiden tunnistamisessa kuin kuuntelukokeeseen osallistuneet kuulijat.

3. Lähdeääni ei heijastellut ainoastaan F0:n ja L_{eq}:n muutoksia, vaan sillä näytti olevan myös itsenäinen rooli emootioilmaisussa. Lähdeäänen laatua kuvaavista parametreistä NAQ erotteli valenssia alfa ratiota selkeämmin molemmilla sukupuolilla.

4. Formanttitaajuudet F1, F2, F3 ja F4 liittyivät valenssin vastaanottoon [a:]-vokaalista molemmilla sukupuolilla. Positiivisen valenssin vastaanotolla näytti olevan yhteyttä korkeaan F3:n taajuuteen.

5. Yhdeltä korkeudelta tuotetutuilla vokaaleilla [a:], [i:] ja [u:] näytti olevan keskenään erilainen kyky välittää emotionaalista informaatiota.

6. Puheeseen liittyvän emootioilmaisun kompleksisuus näkyi suurina yksilöllisinä eroina. Puheen redundanssi sallii puheen parametrien erilaiset keskinäiset suhteet. Parametrien interaktioiden vaikutukset kuulohavaintoon vaatisivat jatkotutkimuksia.

7. Äänilähteen ja ääniväylän vaikutusten välistä suhdetta eri vokaaleissa on syytä tutkia jatkossa synteesin avulla, mikä mahdollistaa halutunasteiset muutokset pelkästään valituissa muuttujissa.

8. Miehet jättivät vastaamatta tilastollisesti merkitsevästi suurempaan määrään esitettyjä näytteitä kuin naiset. Tämä saattaa kertoa miesten suuremmasta epävarmuudesta emotionaalisen informaation vastaanotossa. Sukupuolten välisiä eroja tunnetilojen vastaanottamisessa olisi syytä tutkia tarkemmin, esimerkiksi aivotutkimuksen keinoin.

1. Introduction

1.1. Background of vocal expression

The human voice is an extremely flexible medium and one of the most important means of conveying and exchanging information between people. The specific human capability to produce and perceive vocal speech sounds has required, firstly, an anatomical development of the vocal tract with a wide phonetic range; secondly, an adapted neural development underlying vocal control and imitating vocal sounds; thirdly, a capacity for vocal learning; and fourthly, speech perceptual specializations (Fitch 2000). Commonly, Broca's area in the brain has been seen as the seat of motor control of speech (Rizzolatti and Arbib 1998). According to Rizzolatti and Arbib (1998), however, "the motor properties of human Broca's area do not relate only to speech (…), Broca's area might also become active during the execution of hand or arm movements, during mental imagery of hand grasping movement (…), and during task involving hand-mental rotations". Hence, Rizzolatti and Arbib have suggested that "the precursor of Broca's area was endowed, before speech appearance, with a mechanism for recognizing actions made by others". This suggests a link between action perception and action production, a mirroring neural mechanism in the brain. The discovery of the mirror neurons is seen as a bridge between action representation and the development of inter-individual communication, phonetic gestures and ultimately of speech (Rizzolatti and Arbib 1998). Thus, speech may be perceived rather by recognition of the articulatory movements than the speech sounds alone (Nishitani and Hari 2002). (For the mirror neurons, see e.g. Gallese V, Fadiga L, Fogassi L, Rizzolatti G. Action recognition in the premotor cortex. Brain 1996: 119, 593-609; Rizzolatti G, Fadiga L, Gallese V, Fogassi L. Premotor cortex and the recognition of motor actions. Cognitive Brain Research 1996: 3, 131-141.)

The most important characteristics in human speech are the formants, which is verified by whispered speech and by sinewave speech, both of them based on the acoustic structures of the

formant frequencies (Fitch 2000). The use of the wide formant pattern range is enabled by the relatively lower position of the larynx in humans than in other primates (Lieberman 1969; 1972). This further enables both vertical and horizontal tongue movements. The localization of the lips, teeth, tongue, pharynx and intra-abdominal functions on the somatosensory cortex and the functions of vocalization, salivation and mastication on the motor cortex are bilaterally very close to each other (Chusid 1976, 6). Mastication (chewing, sucking and licking), a continual mouth open-close alternation may form the base of articulatory movements (MacNeilage 1998). The synchronized motions of articulators, the larynx and possibly of other body movements require a sophisticated neural system. Gradually this mandibular oscillation began to assume communicative significance as lip and tongue smacks and teeth chatters (MacNeilage 1998). This "frame" of the mastication movements may also have given the rhythm for the syllables of speech (MacNeilage 1998). The unusual human ability to imitate sounds has been seen as a prerequisite for the formation of the large vocabularies that typify all human languages (Fitch 2000). Vocal imitation in the auditory domain has also been regarded as easier than other forms of imitation since a person can hear his/her own voice, and compare the vocal output with that of others (Fitch 2000). The basis of the ability to imitate sounds may lie on the action of mirror neurons or more likely in the audiovisual mirror neurons. The mirror neuron system provides motor copies of the observed actions while the audiovisual mirror neurons do not code space or spatial characteristics but actions when they are only heard (Kohler et al. 2002). This suggests that the articulatory gestures can be recognized by others even if they do not see the speaker. The audiovisual mirror neurons code abstract content, the meaning of actions, and by so doing, they have auditory access to contents that are characteristic of human language and may thus act as a part of the origin of language (Kohler et al. 2002).

There is also a bio-evolutionary origin in John Ohala's frequency code theory (1997) which suggests a non-arbitrary connection between certain speech sounds and meanings. According to this theory, there are similarities across species in the recognition of "sound symbolism". There

smallness is communicated by high acoustic frequency (non-threatening, submissive) and largeness by low acoustic frequency (threatening, superior) (Ohala 1997). Xu and Chuenwattanapranithi (2007) have further found that there is a link between the estimation of the body size and that of the emotional state; both the estimates seem to be sensitive to vocal tract length and fundamental frequency (F0). Synthesized vowels produced by a lengthened vocal tract and lowered pitch were perceived as from a large person and the opposite set with shorter vocal tract and heightened pitch level as from a smaller person. Dynamic variations in these were distinctive and affected the perception of whether the samples were identified as joy or as anger (Xu and Chuenwattanapranithi 2007; Chuenwattanapranithi et al. 27.10.2008).

Vocal messages tend to be more or less colored by emotional meanings, positive or negative, subtle or strong which, in turn, appear to form a central source of vocal variation. Thus, the voice, not only the linguistic content of the speech, may also act as a powerful messenger of the physiological and psychological state of the speaker. These universally recognizable elements of vocal expressions of emotions appear to be based primarily on evolutionary physiological changes in the body, and tend to be developed in interaction with the survival-environmental events (Darwin (1872) 1934; Damasio 1999; Izard 2007) award to a complex inter-individual psychophysiological form of communication. Changes in vocal expressions caused by emotions were first paid attention to by Darwin (1872, 1934).

In the conveyance of emotions, the functions of voice quality - a combination of voice source and the vocal tract function (Fant 1970) - and its components have been studied far less than prosodic parameters, mainly due to conceptual and methodological difficulties (Scherer 1986). Nevertheless, the importance of voice quality in this respect has been appreciated. For instance, Campbell and Mokhtari (2003) have suggested that voice quality should be considered a prosodic characteristic, along with F0 (fundamental frequency), duration and amplitude since in speech production, voice quality conveys paralinguistic differences in meaning. Gobl and Ní Chasaide

(2003) have also concluded that a verbally neutral expression may evoke widely differing associations due to the changes in voice quality alone. They noted further that when expressing strong emotions, wide pitch variations seem to have an important role, while voice quality seems to be the conveyer of more subtle variations. Similarly, voice quality has also been found to be the main parameter differentiating between *secondary* emotions (Murray and Arnott 1993) (for secondary emotions see 1.2.). The findings of Ladd et al. (1985) suggested that F0 range and voice quality may vary individually in the communication of emotions.

The frequently studied prosodic characteristics F0, SPL (sound pressure level) or L_{eq} (the equivalent sound level) and speech rate are typically computed directly from the output of the human vocal apparatus, the speech pressure waveform. In the present dissertation the two components glottal airflow and vocal tract filter are separated from the speech signal according to Fant's source-tract theory of voice production (Fant 1970). The concept of voice quality is studied as consisting of the voice source characteristics, i.e. (quasi-)periodic airflow pulsation resulting from vocal fold vibration, and vocal tract filter function.

1.2. Affect, emotion and valence

Emotions have been of great interest in many sciences, such as psychology, philosophy, social sciences and communication studies. There are several more or less contentious approaches to the emotion theories, beginning from Plato, Aristotle and especially the Stoic emotion theories of classical antiquity (Nussbaum 2001; Solomon 1998). The recent theories also have different approaches to the concept of emotion, some of them dividing emotions into basic or primary and secondary emotions and others including factors like valence and activity into the research theory. Some theories consider emotions through their hierarchical relationships between members in the same emotion category and between the different emotion categories (Guerrero, Andersen and Trost 1998). In spite of disagreements, it is widely agreed that there are basic or primary emotions which are universal, not culture dependent. Into these primary emotions are often included emotions like joy, anger, fear sadness, disgust and surprise (Murray and Arnott 1993). The primary or basic emotions are considered to represent evolutionary and survival-related patterns of responses to environmental events, and are therefore claimed to be universal. They represent either high activity level (or high arousal) through emotions like joy, anger, fear, disgust and surprise or low activity level through emotions such as sadness. It seems that primary emotions have a strong naturally based connection to nature and man. Thus, it is natural that similar expressions of emotions emerge in different ways, e.g. speaking, facial expressions, movements and musical expression (e.g. Scherer 1995). Richman (2001, p. 301) has stated that "in the beginning speech and musicmaking were one and the same: they were collective, real-time repetitions of formulaic sequences". The capability to repeat is emphasized also by Imberty (2001, p. 449): "Repetition, variation, and rhythm in both games and speech, and cognitive-affective exchange, are at the origin of temporal experiences that predispose human beings toward comprehension and creation of musical activities." Musical composition has benefited these collective axioms in emotional expressions.

A well known example of this is the music composed by Bernard Herrmann in Alfred Hitchcock's *Psycho* (1960). There the tense harsh cutting voice of the monotonous violins is well remembered and recognized as expressing fear. In contrast, we do also have an idea of a melody and rhythm of a hymn (sadness) and how it differs from a Christmas carol (joy). We know this since the emotions they are composed to express are given by nature, not created by human minds and thus, they are universal (Feldman Barrett 2006; Izard 2007). Each basic emotion appears to have properties which distinguish it from other emotional states. However, emotions seldom occur in isolation; instead they tend to emerge in clusters or blends (Guerrero; Andersen and Trost 1998) or in emotion schemas (Izard 2007). It has also been argued that in the evolutionary process emotions occurred earlier than their perceptual process and that these two developed in different ways fairly independently of each other (Izard 2007).

The secondary emotions (e.g. uncertainty, diffidence, love, empathy) are affected by culture, and hence, some of them may not be universally recognizable across world cultures. Moreover, different degrees of emotional expressions form further emotion families or categories (see e.g. Polivy 1981; Murray and Arnott 1993; Scherer 1995; Banse and Scherer 1996; Guerrero et al. 1998; Gobl and Ní Chasaide 2003; Schröder 2003). In order to arrange these categories the different degrees of feelings, affects and emotions have been defined as follows: "affect refers to the general valence of an emotional state, emotion refers to specific types or clusters of feelings that occur in response to particular events, and moods refer to relatively enduring and global states of pleasant or unpleasant feelings" (Guerrero et al. 1998). The definitions of the terms affect and emotion are indeed needed, since their use in the literature is not established. The terms affect and emotion are often used synonymously or they may have been defined as Guerrero et al. suggested above or the other way around. Thus, defined expressions for the categories of feelings are needed in order to be more precise in the exploration of this phenomenon. According to the well-known neuroscientist Antonio Damasio an emotional stimulus first evokes a physiological affective reaction in the human

body (Lat. *afficere/affectum* = state of mind caused by a stimulus, excitement) (Damasio 1999, 2003). Then the impulse is conveyed by the nervous system to the brain where it forges ahead through amygdala and some other parts of the brain to the frontal lobe where it is finally understood as an emotion (Lat. *emotio* = e- = away, *movere*/*motum* = to move). In this complicated process the frontal lobes play the most important role (Luria 1973). Emotion tends to indicate action, a new process and a new motivation (Lat. *movere*). Defined in this way the perception of emotional expressions can be divided into two categories: bio-physiological (affect) and cognitive perception (emotion). Affect appears to refer to a more primitive emotional response to a stimulus. Hence, it appears to need a source. In the short-term response to this source, affect tends to operate on the axis positive (good) – neutral – negative (bad) feeling. Affect occurs unexpectedly, and thus, response tends to be holistic, not sense-specific to the stimulus. Emotion seems to lie on a more cognitive level than affect. It may also occur over a longer time span. In contrast to affect, emotion may act as a controller over an affective reaction by "making the decision" over the reaction to an affective stimulus. Thus, emotion refers to a more discrete mental state than affect. Emotion also seems to be more sense-specific than affect, and thus, interpretation of an emotional stimulus may be sometimes difficult if the visual and auditory information are contradictory. Moreover, intra-individual interpretations also play a role in the identification of the perceived stimulus. There, it is a matter of the individual's personality, memory and the background emotion (see 1.3.). Theoretically, this can be called the Hermeneutic Circle, which is formed by earlier knowledge and interpretations, and hence have an effect on the new ones. This is also called the Hermeneutic Spiral due to its ability to go further. Partly, the definition of the concept of priming comes close to the idea of Hermeneutic Spiral. Priming is defined by Fecteau et al. (2004b) as "a type of implicit memory, a nonconscious influence of past experience on current performance of behavior". However, it should be noted that the term priming has been used in a different meaning in laboratory conditions to refer to the repetition of stimuli given.

An emotion can often be named by several attributes unlike an affect. However, the neural connection between affect and emotion (involuntary and voluntary processing) in the brain is extremely fast as if they happen simultaneously. However, recent brain research has shown their temporal difference (Belin 2004; Bostanov 2004; Wambacq 2004). Finally, mood implies the most general and non-specific feeling which does not have any motivated object. Similarly to affect, mood can be described by its valence on the axis positive – neutral – negative. Valence refers to the affective quality of an emotional expression or the perception of an emotional stimulus whether pleasant, neutral or unpleasant (e.g. Murray and Arnott 1993; Zei Pollermann 2002). Normally, affective states like joy or tenderness are considered positive and e.g. anger and sadness negative. Emotion is more complex than affect or mood. Because of the incoherence of the use of the terms affect, emotion, basic or primary emotions and secondary or social emotions and on the basis of the references above, I propose the following distinction between the terms affect and emotion (**Table 1**). Here "affect" includes the terms "basic emotion" and "primary emotion". The term "emotion" includes the "secondary emotion" and "social emotion".

Table 1. Definitions of the terms affect and emotion and their differences.

Affect	Emotion
- primary (or basic) - older parts of the brain - cannot be taught - biophysical reaction to a stimulus - acute - universal - may lead to emotions - involuntary	- secondary - newer parts of the brain - can be learned - cognitive reaction to the stimulus - short- or long-term - culturally dependent - may create new emotions (clusters or schemas) - motivated, more or less voluntary

In the present dissertation, an attempt was made to identify valences, both expressed and perceived, from the acoustic spectrum by measuring formant frequencies (see Laukkanen et al. 1997) and by calculating NAQ from the inverse filtered signal (see 1.7.) (Alku et al. 2002) and by comparing the results from these parameters to the results of the listening tests.

Since affects hardly ever occur alone and since they are commonly mixed together with social emotions the term that will be used to refer to both of these concepts (affect and emotion) in the present dissertation is `emotion´. Additionally, as the samples studied here are voluntarily produced by the actors the resulting emotional states are called `emotions´ (not `affects´).

1.3. Background emotion, expression and perception

The concept of neutrality in the context of emotions is discussed by Damasio (2007). According to him one can be in a neutral state only when he or she is unconscious. Otherwise we are always in some emotional state which he calls a background emotion. Izard (2007) has evinced a similar idea to Damasio's background emotion. He assumes that there is a continuous emotional state which functions as an organiser of the consciousness and hence, affects one's mind, behaviour and social competence. This ongoing emotion affects the perception of emotional stimuli. For instance, if a person is prone to perceive cues of anger he or she may find those cues even where they do not exist. Thus, perception may be biased and the process may become a complex "emotion-cognition-action system" (Izard 2007; Zei Pollermann and Izdebski 2008). According to Izard (2007) the perceptual process of emotions functions in two ways: "Because emotions determine perceptual selectivity and thus influence the processing of information from particular events and situations, they are both proactive as well as reactive in relation to cognition and behaviour." Zei Pollermann and Izdebski (2008) have suggested that such a cognitive process necessarily evolves in a three-dimensional emotional space formed by valence, arousal and potency. They have stated further that "the subject's perceptual-sensorimotor-conceptual schemata activated for purposes of goal driven action always include affective components" (Zei Pollermann and Izdebski 2008).

Luria (1973) perceives perception as an active process "which includes the search for the most important elements of information, their comparison with each other, the creation of hypothesis concerning the meaning of the information as a whole, and the verification of this hypothesis by comparing it with the original features of the object perceived". This process described by Luria is selective and also directive organized behaviour, and thus determines where one's attention is focused. This feature distinguishes the process from a general basic "arousal reaction" since the

origin of a voluntary attention is not biological but social, developed in a social environment (Luria 1973). As emotion is defined the way described here, it cannot be considered the opposite to reason. In fact, emotion as a cognitive process is a cause and a consequence of reason. If there were no reason, there would not be any emotion. Again, Izard (2007) says: “in the normal mind, cognition does not occur in the absence of all emotion”.

1.4. Emosphere

In order to create a methodological system in emotion research it is suggested here that the terms affect and emotion are to be defined as presented above (see 1.2. and 1.3.). Further, these terms (affect and emotion) as well as background emotion could be included in a wider concept which would cover the whole emotional field. Within this field the terminology is defined and hence, the mixed use of the terms would be avoided. This field could be called emosphere, in accordance with Juri Lotman's (2005) concept of semiosphere which he first defined in 1984 in an article written originally in Russian (Sign System Studies 1984: 17, 5-23) (Lotman 2005). Given Lotman's definition of semiosphere, emosphere could be based on the idea that emotions cannot be born or do not exist in isolation. They are a necessity to human life. From the phylogeny viewpoint an individual is born with affects. The socialisation process of the individual differs in different cultures, and the intra-individual development is always unique. The individual's empirical world in interaction with biogenetic heredity is processed in the Hermeneutic Spiral (see 1.2.). Hence, the emotional development of an individual is influenced both by universal and cultural and by intra- and interpersonal aspects. This continuum of the emotional development is called emosphere. I propose a description for the emosphere in a four-dimensional field (**Figure 1**).

Figure 1. Four-dimensional emosphere. Rows: intra- and interpersonal aspects; columns: universal and cultural aspects.

Emosphere

Universal Intrapersonal	Universal Interpersonal
Cultural Intrapersonal	Cultural Interpersonal

As the semiosphere is by nature abstract, so also is the emosphere. Emosphere is connected both to homogeneity and individuality, external and internal (see Lotman 2005) with continuous interplay between and over the boundaries (see **Figure 1**). This vivid action of the emosphere mirrors all the emotional processes ongoing in our minds, some of them being longer and some shorter in duration, and some of them being stronger and some lighter in the impression they make. This mixture is never affected by only one emotion. There are always simultaneous emotional processes, however, the emosphere may be dominated by a single emotional state, e.g. good or bad mood, or sorrow or happiness. Hence, the emosphere is in a continuous process of interpreting, choosing, forgetting and remembering. This dialogue may change the core structures of the emosphere. For example, a markedly negative experience making a strong impression on the individual's mental health is prone to change the personality (intrapersonal) and also his/her social relationships (interpersonal). A similar chain of events could also be initiated by a positive emotional experience. Hence, the dialogue also creates new meanings of the emotional information. According to Lotman in the semiosphere of language, there is no meaning without communication; dialogue precedes language and gives birth to it (2005, 228). Similarly, in the emosphere, there do not seem to be affects, emotions or anything emotive without a meaning. Affects tend to have an important evolution-related meaning and social emotional states seem to be built on the consciousness (see 1.3.) which in turn, provides meanings. The emospheric dialogue is communication between and over its boundaries (**Figure 1**) and hence, it carries meanings, i.e. functional emotional information.

I propose the use of the new term emosphere to conceptualise the multiplicity and discrepancy of the emotional development, experience and behaviour of an individual or a (sub)culture. The awareness of the emosphere may explain and clarify occurring difficulties and misunderstandings and it may also help to understand the reasons behind the differences in the emotional communication between individuals and different cultures.

1.5. Psychophysiological activity level

There are some common characteristics through which all emotional expressions can be observed, e.g. variations in sound pressure level, articulatory movements or changes in the tempo of the heartbeat (Feldman Barrett 2006). Prosodic variables like F0, SPL, temporal aspects such as word – pause relation, duration of a phoneme or a syllable have largely been studied in relation to emotional expressions (e.g. Lieberman 1962; Murray and Arnott 1993; Mozziconazzi 1998; Scherer 2003). Most of all, F0 and its variations have been considered the most important acoustic parameters in the detection of emotional relevance. Murray and Arnott (1993) have stated that "the pitch envelope (i.e., the level, range, shape, and timing of the pitch contour) is the most important parameter in differentiating between basic emotions". Leinonen et al. (1997) have also reported similar results. F0 may vary quite widely and it may also be dependent on the individual's own characteristics or qualities, such as range. Therefore, no special patterns may be needed as transmitters of the affective content of the speech. However, some intonation patterns may simply carry the emotional information better than others (Mozziconazzi 1998).

In importance, changes in F0, which are heard as pitch variations, are most likely followed by energy, duration and speech rate (see e.g. ten Bosch 2003). Additionally to F0, SPL can be regarded as a distinguishing parameter between emotional expressions. In a number of investigations, F0 as well as SPL have been reported to increase in emotions with high arousal, e.g. anger, and to decrease in low arousal affects, e.g. sadness when compared to a neutral emotional state (Banse and Scherer 1996; Laukkanen et al. 1997; Leinonen et. al. 1997; Toivanen et al. 2008). Along with high F0 and intensity, a large number of accents have been reported in the aroused states of happiness and anger (Makarova and Petrushin 2003).

Additionally to valence (see 1.2.), the level of psychophysiological activity is a distinctive coder of arousal of affective or emotional expression. It operates on the axis low – medium – high

psychophysiological activity level. Conventionally, high activity level and thus high muscle activity and hyperfunctional phonation type is connected to emotional states like anger and joy, and low activity level, low muscle activity and hypofunctional phonation type to emotions like tenderness and sadness. Activity level is studied in the present dissertation by calculating the normalized amplitude quotient (NAQ), which is a voice quality parameter related to the phonation type (Alku et al. 2002), and alpha ratio, which is the difference in the level between the upper and lower frequency ranges in the spectrum (Frøkjær-Jensen and Prytz 1973) (for NAQ see 1.6. and for the alpha ratio 3.2.).

On the one hand, glottal characteristics are known to co-vary with F0 and SPL, on the other hand, there are some earlier investigations which have studied the individual role of acoustic voice quality in the expression and perception of emotions and their valence (e.g. Laukkanen 1996). The results of Laukkanen et al. (1996) suggest that voice source parameters may also vary independently of F0 and intensity in emotional utterances. In an other study by Laukkanen et al. (1997), the variation of F0, intensity and duration were artificially eliminated. The samples of short duration (200 ms) used in the study appeared to be categorized by the listeners according to the psycho-physiological activity level inherent in the emotions. The vocal effort level tended to be related to the glottal voice source wave form. Perception of valence seemed to be related to the first formant (F1) and to the forth formant (F4) (Laukkanen 1997). The results from a subsequent study by Waaramaa et al. (2007) were quite similar. There the role of F0 was eliminated as the subjects, student actors, were asked to produce different emotional states expressed in mono-pitched vowels [a:, o:, e:]. Intensity was allowed to vary freely in the stimuli, however, SPL was normalized for the perceptual analysis. The glottal voice source characteristics reflected in alpha ratio and NAQ were the two parameters to which the perception of the psychophysiological activity level of the heard emotions tended to be mostly related. The results also implied that the formant frequencies may be of relevance in valence perception, especially formant frequencies F3 and F4 having higher values

in positive emotions than in negative ones (Laukkanen et al. 1997; Waaramaa et al. 2007; Laukkanen et al. 2008).

1.6. Voice quality

Psychophysiological activity level is related both to the valence of an affect and to its degree of strength, which in turn is closely related to voice quality (Murray and Arnott 1993). Gobl and Ní Chasaide (2003) have classified high activation or arousal in a tense/harsh voice quality and, in turn, low activation in a breathy or creaky voice quality. Consequently, the variations in arousal influence the listeners' judgments of the emotional content of the speech (Ladd et al. 1985). Thus, voice has a strong impact on the impressions the speaker gives of him/herself, how the message is perceived or even understood, and what kind of (emotional) feedback it evokes in the listeners (Lieberman 1962, Addington 1968, Weaver and Anderson 1973, Blood et al. 1979). Emotional information perceived by more than one sense gives a more reliable conception of the substance of the message than if the information is perceived only by one sense (Van den Stock et al. 2007). Vocal stimuli have considerable individual variance and hence, their interpretation by one sense (hearing) alone is complicated, especially in the absence of familiarity. It has been shown that familiarity in the speaker's voice enhances its auditory processing in the listener's brain (Birkett et al. 2007).

Laver (1980) defines voice quality in a broad sense, as "the characteristic auditory coloring of an individual speaker's voice". In a narrow sense, voice quality may be either a short-term or a long-term feature. In the first case voice quality may refer to any single vocal characteristic (e.g. pitch or loudness of the voice) signalling for instance, the speaker's current emotional state. In the second case voice quality may differentiate between speakers or groups of speakers, for instance social or cultural groups.

Voice quality is a combination of two factors, the voice source and vocal tract function (Fant 1970). Voice source is formed by air flow modulated by vocal fold vibration; it determines the fundamental frequency and has an effect on SPL or L_{eq}. (The main difference between SPL and L_{eq}

is that SPL measures the loudest dB peaks of the sample while L_{eq} takes into account the whole duration of the sample and measures the mean of dB energy level in it.) Vocal tract function is formed by vocal tract resonances, i.e. formants. Both of these functions are seen in the manner sound energy is distributed along the frequency range in the spectrum. In a more hyperfunctional phonation type the spectral slope is flatter than in a hypofunctional phonation type, where the slope is more tilted (Gauffin and Sundberg 1989). Phonation type refers to the relation between the subglottal air pressure and the amount of adduction in the glottis. These actions determine the speed of the collision of the vocal folds. A rapid collision of the vocal folds and complete closure of the glottis result in stronger overtones and hence, in a flatter spectral slope. For slower closing and incomplete closure the overtones are weaker and their perceptual relevance is smaller. Here the slope is also more tilted (Gauffin and Sundberg 1989). These phonation types (hyperfunctional and hypofunctional) are perceived respectively as pressed or breathy voice qualities.

A connection has been detected between voice quality and duration of speech: vowels produced with a `breathy´ (hypofunctional) voice quality have been reported to have longer duration than those produced with a more strained (more hyperfunctional) voice quality (Wayland and Jongman 2003). In some cases, temporal aspects tend to be used as distinguishing features in the conveyance of emotional content in speech. The expressions of anger, disgust, fear and shame appear to be differentiated from emotions like sadness, joy and guilt by, among other things, their shorter duration (Scherer and Wallbott 1994).

Bostanov and Kotchoubey (2004) have stated: “Clearly, the shorter an emotional utterance (verbal or nonverbal), the fewer prosodic features are present in the narrow sense of the term [prosody] and the more important the voice quality is for affect recognition. Thus, it is reasonable to hypothesize that short exclamations convey emotion predominantly through voice quality.” Short vowel samples (~ 100 – 2500 ms) were observed in order to exclude possible speech prosodic effects and to reveal the voice quality characteristics more clearly. The use of samples as short as

< 200 ms proved adequate, since, according to neurological findings on brain responses to emotional stimuli, the recognition of emotional information takes place within the first 30-160 ms of the expression (Bostanov and Kotchoubey 2004; Damasio 2003; Izard 2007; Zei Pollermann 2002). This recognition appears to be based primarily on voice quality (Bostanov and Kotchoubey 2004). These findings concurred with earlier results by Pollack (1960). Additionally, a greater activity of the cortex has been observed when attention has been directed to the speaker's voice compared with the verbal content (Belin et al. 2004, 132). This may be due to fact that the processing of the human voice occurs in an auditory domain in the left human prefrontal cortex distinct from Broca's area, which, in turn, processes syntactic and semantic actions (Fecteau et al. 2005b, 2253-2254). Moreover, the auditory cortex is species-specific to vocalizations: nonhuman or other types of artificial stimuli or animal vocalizations do not evoke as strong responses in this area of the brain as do human sounds (Fecteau et al. 2004a). However, a simple affect burst does not last long enough to convey an expression of an emotion (ten Bosch 2003), and therefore, the samples examined here were gathered from continuous speech and prolonged vowels.

Generally, the material used in vocal emotion research has been gathered from sentence-length or longer passages of texts (Addington 1968; Weaver and Anderson 1973; Blood et al.1979; Laukkanen et al. 1996), including automatic classification experiments (e.g. McGilloway et al. 2000; Yu et al. 2001). In the present dissertation, the focus was on shorter phoneme-length utterances. This seemed reasonable since it was of interest to study voice quality, not the degree of "correctly" recognized, identified or discriminated emotion samples.

As the three parameters F0, SPL and duration, are the most obvious carriers of emotional information the potential role of voice quality *per se* was investigated. In the present dissertation, the concept of voice quality may be distinguished from the term voice timbre. Here the term voice quality also extends to glottal aspects like phonation type, turbulence noise and short-term temporal characteristics (perturbation). The timbre of the voice refers mainly to its coloring on an axis

brightness – darkness. Timbre is said to be an attribute characterizing a given voice and distinguishing it from other voices when the pitch and the vowel are identical (Cleveland 1976; Sundberg 1977). Timbre has also been defined as “a subjective aspect of sound for which there is no such scale and neither qualitative nor quantitative descriptions generally found that are widely accepted” (Howard and Tyrell 1997). Thus, the term voice quality is used here.

1.7. Inverse filtering and time domain parameterizations of the glottal flow

In order to study voice quality it is crucial to examine the phonation type. One way is to study the pulse shape of the phonation, which is done here by inverse filtering of the voice signal. The voice signal is created by the air flow from the lungs. According to Bernoulli's principle the air flow makes the vocal folds vibrate because of the varying air pressure (Daniel Bernoulli, 1700-1782). The air pressure from the lungs (subglottal pressure) pushes the vocal folds open and the air begins to flow through the glottis. The velocity of the flow is highest in the narrowest parts, with the consequence that air pressure decreases between the vocal folds and sucks them together again. This continuing pressure variation, inversely varying velocity and the elasticity of the vocal folds make the vocal folds vibrate. This vibration forms the voice source (see 1.5.). The vibration of the vocal folds can be studied by e.g. photoglottography or high-speed digital image recording system and the glottal contact area variation by EGG (electroglottogarphy) technique.

It is possible to investigate the glottal volume velocity waveform by inverse filtering methods. The inverse filtering method is based on the Fant's (1970) theory of voice production, which assumes that the human voice production system consists of two parts: firstly, the voice source and secondly, the vocal tract. Additionally, there is a third part which is of importance in inverse filtering (of the acoustic speech sound): the lip radiation effect. By cancelling out the estimated effect of the vocal tract function (formant frequencies) and the lip radiation the study of the glottal pulse form (voice source) becomes possible (Alku 1991).

Inverse filtering can be performed either based on the airflow (e.g. Rothenberg 1973) or on the acoustic speech pressure signal, e.g. IAIF (Iterative Adaptive Inverse Filtering) method (Alku 1991; 1992; Vintturi 2001). In Rothenberg's pneumotachograph mask method the original glottal waveform is picked from the oral airflow. In this method, the use of the mask affects speech production. IAIF is one of the methods based on the inverse filtering of the acoustic speech pressure

signal and hence is non-invasive. The IAIF method was used in the present dissertation since it does not impede the speech production or affect the outcoming acoustic signal as the mask does by lengthening the vocal tract and thus altering both the natural voice production and the resulting signal (Alku et al. 1998b). Furthermore, IAIF is an adequate method especially in studying higher frequencies up to 4 kHz, whereas the pneumotachomask frequency response limits the study to about 1.5 kHz (Rothenberg 1973; Alku et al. 1998b).

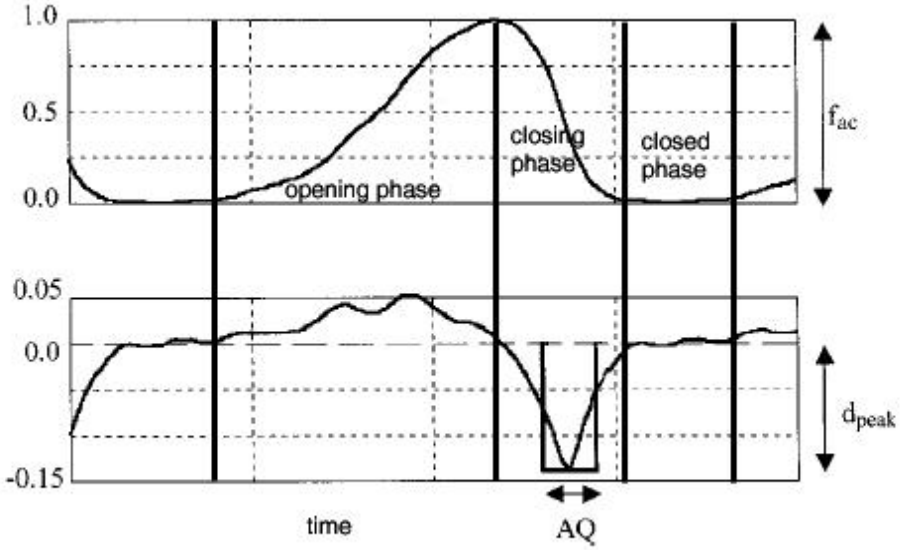
The performance of the inverse filtering depends largely on how accurately the formant frequencies can be estimated. The estimation of the formants becomes difficult when a voice has a sparse harmonic structure. The accuracy typically deteriorates for samples with high F0, such as female voices.

The glottal signal can be characterized by various time based parameters: open quotient (OQ), the open time of the glottis divided by the period length, closed quotient (CQ), the closed time divided by the period length, closing quotient (ClQ), the closing time divided by the period length, and speed quotient (SQ), the opening time divided by the closing time of the glottis. Raising F0 typically increases SPL by increasing the glottal closing speed/speed of flow cessation in a pulse, and consequently decreasing OQ. However, raising F0 may also increase OQ, especially in falsetto. In general, SQ and CQ increase, and ClQ decreases with higher intensity (Fant 1987; Holmberg et al. 1988; Gauffin and Sundberg 1989; Granström and Nord 1991; Laukkanen 1995; Alku et al. 1998a).

As the phonation type also has an effect on the pulse shape (Gauffin and Sundberg 1989) amplitude based parameterizations of the glottal waveform were developed. Normalized amplitude quotient (NAQ) is one of them. Amplitude quotient (AQ) (**Figure 2**) is a time-domain quantity which correlates strongly with the length of the glottal closing phase but is shorter than the true closing phase (Alku et al. 2006). NAQ is the normalized version of AQ. NAQ is determined as a ratio of the flow amplitude (the peak-to-peak AC flow (f_{ac})) to the negative peak of the first

derivative of the flow waveform (d_{peak}), multiplied by the period length (T) (NAQ = $f_{ac}/(d_{peak}T)$) (Alku et al. 2002). NAQ is measured from the estimated flow inverse filtered e.g. with the IAIF method (Alku et al. 1992). NAQ has been shown to be sensitive to different phonation types. Alku et al. (2002) measured a smaller NAQ in a hyperfunctional voice where the glottal closing phase was shorter and the closed time longer than in a hypofunctional voice quality. A strong correlation has also been reported between SPL (dB) and NAQ (Alku et al. 2002). NAQ is more robust to distortion such as formant ripple compared to e.g. ClQ (Alku et al. 2006).

Figure 2. Model of inverse filtered signal. Upper figure: Glottal flow; f_{ac} = AC flow amplitude. Lower figure: First derivative of glottal flow; d_{peak} = negative peak amplitude of the derivative. AQ (amplitude quotient): f_{ac}/d_{peak}. Time is on the horizontal axis, flow on an arbitrary scale on the vertical axis.



The shape of the vocal tract is determined by articulation (whether lengthened or narrowed). These modifications change the transfer function of the vocal tract. The impedance increases as a result of lengthening or narrowing of the vocal tract. Then more subglottal pressure is needed for a certain amount of air to pass through. The relation of the frequencies between F0 and F1 affects the

pulse shape. The closer F0 and F1 come to each other the more the skewing of the pulse shape increases. The skewing of the glottal pulse seems to be a result firstly of the phonation type and secondly of the inertia effect of the vocal tract. When the frequencies of F0 and F1 are equal the pulse shape becomes more symmetrical. Skewing increases the flow cessation rate and strengthens the partials. Hence, SPL tends to increase somewhat. The voice becomes closer to its modal register and brighter and more metallic in quality. A more symmetrical pulse is related to a slower flow cessation rate and a steeper spectral shape and the voice sounds more like falsetto. This relation between the vocal tract and the pulse shape is called acoustic interaction (Fant 1970; Rothenberg 1981; Sundberg and Gauffin 1978; Gauffin and Sundberg 1989; Cummings and Clementes 1995; Story et al. 2000; Titze 2008). F0 *per se* may have an effect on vocal intensity in a loud type of phonation; in soft and normal phonation the speakers may control their vocal intensity by the glottal flow shape and amplitude (Fant 1970; Alku et al. 2006; see also Vintturi 2001).

1.8. Articulatory characteristics

In addition to the voice source characteristics, voice quality is also affected by articulatory characteristics. Articulatory characteristics may vary in relation to prosodic variables. For example, raising the F0 normally leads to an elevated laryngeal position. Hence, a higher F0 may lead to a shortened vocal tract with higher formant frequencies (Holmberg et al. 1989; Shipp and Izdebski 1975). The first two formants F1 and F2 are of importance in carrying the linguistic task in speech. The amplitude level of F3 has also been found to affect the intelligibility of /y/ (Aaltonen 1985, 1997). Otherwise, the higher formant frequencies F3 and F4 carry more expressive characteristics of the voice. In smiling, for instance, the formant frequencies attain higher positions due to the shortened vocal tract. In earlier studies, the formants F2, F3 and F4 have been found to be higher, especially in positive valence compared to negative valence (Tartter and Braun 1994; Laukkanen et al. 1997; Waaramaa et al. 2007). Higher F1 may also imply a wider mouth opening, which increases SPL by moving F1 closer to F2 (Fant 1970). Wider mouth opening yields higher amplitudes of the formant frequencies overall (Fant 1970; see also Tom et al. 2001). The closeness of F1 and F2, in turn, raises L_{eq} since it increases the amplitudes of the formants (Fant 1970; see also Story et al. 2000). The distance shortened by half between the formants strengthens their peak amplitudes by 6 dB and the frequencies between them by 12 dB (Fant 1970). The perceptual value of a formant is influenced by its amplitude and the sensitivity of hearing in this particular frequency range where the formant frequency occurs as a peak in a spectrum of a speech signal.

Formant amplitude is affected by the strength of the partials, by the tilting of the spectrum, not only by the distance of the formants from each other but also by the formant tuning, how close a match there is between a voice source partial and a formant. The stronger the overtones are in the voice source spectrum the higher are the formant amplitudes. Formants are also sensitive to the widening and narrowing of the vocal tract. Their frequencies change in accordance to these

transverse movements of the vocal tract. A narrowing near the glottis heightens the first formant in frequency whereas narrowing in an upper part of the vocal tract lowers it (Fant 1970, see also Švancara et al. 2006). Generally, the change is greatest if it is near the antinodes of the maximum pressure of the formant in the vocal tract (Švancara et al. 2006). As the articulatory movements have an effect on the formant frequencies, it is plausible that the expression-related acoustic changes are different when occurring in the context of a different vowel. Conversely, the same articulatory movements cause different acoustic changes in different vowels (Fant 1970).

1.9. Identification and computer classification of the samples

In earlier studies emotions have been recognized or distinguished with 15 to 100% accuracy, however, the result has appeared to be dependent on the expressed emotion itself, prosody and voice quality (Polivy, 1981, Banse and Scherer, 1996; Laukkanen et al. 1997; Schröder, 2003). Studies of the perception of emotions have also faced some criticism in the literature concerning the material chosen for the listening tests (Aubergé and Cathiar 2003; Feldman Barrett 2006). The critics claim that the results of the tests may be distorted since the samples are chosen so carefully beforehand by the researchers that only "the best" (or the most conventional) samples have been used in the final studies. In the present dissertation (*Articles I-IV*) the samples were chosen either randomly or all of the produced samples were analyzed and used in the listening tests. Both student actors and professional actors served as subjects. Hence, it has to be taken into account that the skills of the subjects in expressing emotions may vary. The subjects may also vary in their inclinations, some are perhaps more skillful in kinesthetic and facial expressions (visual) while others may be more talented in vocal (auditive) emotional production of emotional expressions. Thus, not all of the samples produced may always have been perceived as intended to represent a certain emotion.

Recently, the interest in emotion studies has grown significantly, including in the new information technology. The prospects this may offer are bringing speech research into a new era. Already quite lot of research of the automatic classification of emotions has been done, and with promising results. Computer algorithms can make automatic classifications and recognitions from large material corpora. They can easily combine different parameters, they can help with speech and speaker recognitions, they have forensic and security applications, also applications for disabled people etc. Research on automatic classification may show correlations and underlying links which could not be observed by other means. Seppänen et al. (2003) have suggested fairly high

recognition percentages for the correct classification of emotional speech samples by automatic discrimination (60-70 %). The best results were obtained by a speaker-dependent classifier. According to Seppänen et al. (2003) the methods to extract prosodic data automatically are still in the development process. So far, most of the research has concentrated on F0 and intensity detection but recently also on voice source parameters. New algorithms and measurement techniques have been developed. The most widely used paradigm in speech recognition systems is the Hidden Markov Model (HMM) (e.g. Lee et al. 2004). HMM is a statistical model for a temporal pattern recognition, e.g. of a speech signal. Another frequently used classifier is the kNN (k = (e.g.) 1, 3, 5; NN = Nearest Neighbor) method (e.g. Seppänen et al. 2003; Morrison et al. 2007). The kNN is a non-parametric method including feature vectors with class information. The classifier compares the unknown vectors to all vector prototypes, picks up the k nearest vector and determines the class of the unknown vector according to the majority of the classes among the k Nearest Neighbors. The kNN classifier was also used in the present dissertation (*Article II*).

1.10. Aim of the study

The aim of the present dissertation was to study phoneme-length vocal units and their voice quality in the expression and perception of emotions. The assumption was that the duration of a single phoneme unit is sufficient for carrying emotional information of speech. However, because of the short duration of the samples it was expected that the number of the "correctly" recognized samples might not be very high. Therefore one of the studies concentrated on perception, asking what kind of voice quality is perceived as a certain emotional state. The interest was to see if there are voice quality parameters which may affect the perception of emotional valence and psycho-physiological activity level other than those frequently studied speech prosodic characteristics, F0, SPL or L_{eq} and duration.

The first study (*Article I*) concentrated on the perception of short emotional samples. In the second study (*Article II*) the aim was to investigate whether there were differences between human listeners and computer classification of the emotional stimuli and what kind of differences they might be. The aim here was also to see how the automatic classifier managed to classify the acoustic features from human phoneme length emotional samples. The third study (*Article III*) concerned semi-synthesized vowels with F3 modifications and the last fourth investigation (*Article IV*) focused on the mono-pitched expressions in different vowels. In the last study the idea was to eliminate the known vocal emotional effect (F0) and to contemplate the plain voice quality.

In order to find out what kind of role, if any, voice quality plays in emotional communication, the effect of pitch variation was eliminated by using short samples (~ 100 – 2, 500 ms) in every study of this dissertation. This strict definition for the research object seemed justified since the technical equipment used in speech and speaker recognition and other applications (e.g. applications for disabled people) are developing fast and more detailed knowledge of ever smaller units is needed in order to create more natural sound quality. The results of the present study may also be

used as basic knowledge for emotional voice production in the education of vocologists, speech communication researchers and actors.

2. Materials and methods

Article I. The speech data for the first study consisted of repetitions of a Finnish prose text read by nine professional actors (N = 5) and actresses (N = 4), aged 26-45, without any known pathologies of the larynx or hearing. The subjects were asked to express the following emotional states while reading: sadness, joy, anger, tenderness and a neutral emotional state. The emotions were produced in random order, with ten repetitions of each emotional state. No detailed instructions were given to the subjects concerning the expressions, e.g. the grade or quality of the emotions like cold or hot anger, or depressed or grief like sadness. The duration of the prose passage read aloud was approximately one minute (in a neutral emotional state). A total of 450 samples (ten samples for five emotions by nine speakers) were recorded in an anechoic chamber using a Bruel & Kjaer 4188 microphone and a Sony DTC-690 DAT recorder. The distance between the subject's mouth and the microphone was 50 cm. The emotions of sadness, joy, anger, tenderness and a neutral emotional state were chosen since they represent both positive and negative valence and high and low psycho-physiological activity levels. From a Finnish sentence "Taakkahan se vain on." ("It is indeed a burden only."), a word [ta:k:ahan] ("indeed a burden") was edited from the prose text read aloud, and the stress-carrying first long vowel [a:] was extracted for further analyses and for a listening test. (Usually, the stress-carrying syllable has a higher intensity than the other syllables (Fujimura et al. 1995)). The SoundSwell computer program was used. It was considered that the listeners would be able to do the test without getting too tired while listening to 200 samples. Thus, 200 samples were randomly chosen for the evaluation. Those samples with a harmonic distortion and those with such a weak or irregular signal that the analysis software could not detect periods were rejected. A total of 171 samples were chosen for the final listening test and analyses.

Article II. The data for the second article was derived from the first experiment. Here the

correct recognitions of the emotion samples were considered, firstly by humans and secondly by the kNN based automatic classifier (see 1.9.).

Article III. The material for the third experiment was derived from the first study recordings. The best recognized [a:] samples of sadness, tenderness and anger (N = 3) were chosen as the material of the study. The perceptual role of F3 in [a:] vowels was examined by applying synthesis for a hypofunctional and hyperfunctional voice qualities in the samples representing tenderness, sadness and anger. The third formant (F3) obtained by the inverse filtering method in the all-pole vocal tract model was modified by both raising and lowering its value by 30 %. Additionally, two more vocal tract settings were computed, one with the original F3 value and one with F3 completely eliminated. Thus, for each of the three emotional states there were four variations of the vocal tract settings in the [a:] vowel: the original, one with F3 30 % higher than in the original, one with F3 30 % lower, and one with F3 completely removed. The desired speech samples were generated through the modified vocal tract filters.

Article IV. Thirteen student actors (5 males and 8 females) with normal voices served as subjects for the fourth study. The subjects produced three mono-pitched prolonged vowels [a:], [i:] and [u:], expressing five emotional states in random order. The samples represent both high and low activity level and positive and negative emotional valences as in the earlier studies: anger, joy, sadness, tenderness and a neutral emotional state. The material was recorded in a well-damped studio using a digital recorder Tascam DA-20 and a Brüel & Kjær 4165 microphone. The distance of the microphone was 40 cm from the subject's mouth. Intensity and duration of the expressions were allowed to vary freely, but the pitch was standardized in order to eliminate the effect of F0 variation and to concentrate on the possible voice quality variations. The resulting material contained 195 mono-pitched vowel samples (13 actors x 5 emotional states x 3 vowels).

3. Analyses

3.1. Perceptual analyses

Article I. The emotionally expressed 171 [a:] vowel samples were presented to 50 listeners (university students and teachers, 41 females, 9 males, mean age 28.5). Their task was to note which emotion they heard. The five emotions expressed joy, anger, tenderness, sadness and a neutral emotional state were the given options in the questionnaire used. The emotional valence was classified by the authors from the responses given. The listening test was performed in a well-damped studio using a digital recorder (Tascam DA-20) and a high-quality loudspeaker (Genelec Biamp 1019 A) from which the listeners' distance was ca 2.5 meters. Each of the 171 samples were replayed to the listeners only once at normal conversational loudness. The number of the samples produced by males was 99 and by females 72. The duration of each sample was ~ 150 ms. The test took 30 minutes.

Article II. The results of the listening test of the first experiment were computed for the second article. A confusion matrix of the expressed and perceived emotional states was formed. Additionally, the same 171 samples were computed using automatic classification methods. The acoustic parameters were used as dimensions and kNN (k = 1, 3, 5) method as the classifier. In order to find out in which class most of the k nearest prototypes belonged majority voting was performed (Toivanen et al. 2006). This determined the class of the unknown feature vector. Leave-one-out testing was used to evaluate the whole data.

Article III. The 12 semi-synthetic [a:] samples were replayed to 30 listeners (24 females, 6 males, mean age 35 years). The Judge computer program (developed by Svante Granqvist, Kungliga Tekniska Högskolan (KTH), Stockholm) replayed the samples at normal conversational loudness in randomized order to each of the listeners. Six samples were repeated in order to study intrarater reliability. The listeners used Sennheiser HD 530 II headphones. The task was to evaluate

the level of emotional valence in the samples, i.e. the positivity, neutrality of negativity of the expressions. The answers were given on a Visual Analog Scale (VAS) on an axis positive – neutral – negative (0 – 1000 units, the neutral point was at 500 units). The listeners were able to have the samples repeated as many times as they wanted.

Article IV. The 195 (N) [a:], [i:] and [u:] vowel samples produced were replayed to 40 university students and teachers as listeners. In order to find out whether there were any gender differences in the listening test results, equal numbers of male (N = 20) and female (N = 20) listeners were recruited for the test. The mean age of the listeners was 38 years (SD = 15) in females and 39 (SD = 11) in males. Two of the female listeners were non-native Finnish speakers; they were exchange students from the Czech Republic. The Judge computer program was used to evaluate the samples. The program replayed the samples in different randomized order for each listener. The listeners could listen to the samples as many times as they wanted but it was recommended by the author to replay every sample only once if possible in order to catch the very first reaction to the stimulus. The intrarater reliability needed to be studied and therefore 15 samples were repeated. The participants did the test one by one using Sennheiser HD 530 II headphones. The average duration of one sample was 2336 ms (SD 1379) in males and 2472 ms (SD 1534) in females. The listeners' task was to identify the emotions the samples were representing. They answered on a Visual Analog Scale (VAS, 0 – 1000 units). The `0´ end was labelled `neutral´ and the `1000´ end was labelled by the emotion produced, i.e. joy, anger, tenderness, sadness or a neutral emotional state.

3.2. Acoustic analyses

Article I. The LTAS (long-term-average spectra) and spectrograms were made from each sample with a signal analysis system called Intelligent Speech Analyser (ISA), developed by Raimo Toivonen M.Sc.Eng. F0 (Hz), L_{eq} (dB) and duration (ms) of the samples were also measured. Alpha ratio (dB) was calculated in order to quantify the spectral energy distribution which reflects the phonation type. Thus, the voice quality along the axis hypofunctional – hyperfunctional perceptually corresponds to the qualities `breathy´ and `strained´. The four lowest formant frequencies F1, F2, F3 and F4 (Hz) were measured in the middle of each vowel from LTAS and spectrograms. The [a:] vowels were analyzed by separating the signal into the glottal flow and the vocal tract function using the IAIF method (Alku 1992). The estimated glottal waveform (voice source) was expressed on an arbitrary amplitude scale. The glottal waveform was parametrized by calculating NAQ.

Article II. The same material was used in the second study as in the first one. In addition to the above-mentioned acoustic parameters (*Article I*), S/N ratio (signal-to-noise ratio), shimmer and the standard deviation and the average of jitter were calculated of the samples for the automatic computer analysis. Jitter describes the cycle-to-cycle variation in a vocal fold period length (Hz). Shimmer is the acoustic parameter to measure the period to period variation in period amplitude (dB). S/N ratio describes the relation between the average harmonic sound energy and the disharmonic sound energy, i.e. noise in the signal. These three variables were not used in the human evaluation tests, since according to preliminary tests, they did not show any statistically significant differences between the emotional expressions.

Article III. As the material for the third study was derived from the first one, the same acoustic analyses also concern the samples used here. In the listening test for the first study, the three best recognized samples of sadness, tenderness and anger were chosen for the material of the third study. These three samples represented both positive and negative valences and both hypofunctional

(sadness and tenderness) and hyperfunctional (anger) voice qualities. The original vowel samples were synthesised in order to obtain samples with different values of F3, one with value raised and one with value lowered value by 30 % and one with F3 totally removed. The modification resulted into four variants: the original vocal tract obtained by inverse filtering, that with F3 30 % higher than in the original, that with F3 30 % lower than in the original and that with F3 completely removed **(Figures 3a-d**). The formant amplitudes were preserved in the modification of the vocal tract filters when computing. The obtained glottal flows were filtered in order to generate the synthetic [a:] vowels in each emotion. The resulting 12 samples were normalized at 70 dB for the listening test.

Article IV. In the fourth study, F0, L_{eq} and duration were measured and spectrograms and normalized average spectra were calculated using FFT (Fast-Fourier Transform). The baseline in the FFT is the maximal spectral peak which is dragged down to a zero value and all other spectral values are compared to it. Formant frequencies F1, F2, F3 and F4 were measured from the peaks in the average spectra and from the darker lines in the spectrograms (300 Hz bandwidth). The analyses were made by ISA. As the actual Hz is naturally connected to the psychophysiological activity level of the speaker and may co-vary with the intensity, F0 was measured to confirm the monopitched expressions and to see possible subtle differences which may reflect the general arousal level. Alpha ratio was calculated. It appeared to be reasonable to calculate NAQ only for the vowel [a:] since the first formant in [i:] and [u:] are so close to F0 or to its second lowest harmonic that it was difficult to reliably distinguish between the source and filter characteristics in the inverse filtering process.

Figures 3a-d. Average spectra of vowel [a:] showing F3 modifications of a tenderness sample originally produced by a male actor. Horizontal upper axis: Frequency in kHz (horizontal lower axis: Bark scale); vertical axis: Relative amplitude in dB. **Figure 3a:** the original sample; **Figure 3b:** the original F3 raised by 30 %; **Figure 3c**: F3 lowered by 30 %; **Figure 3d**: F3 completely removed by filtering. The figures were made by ISA.

Figure 3a.

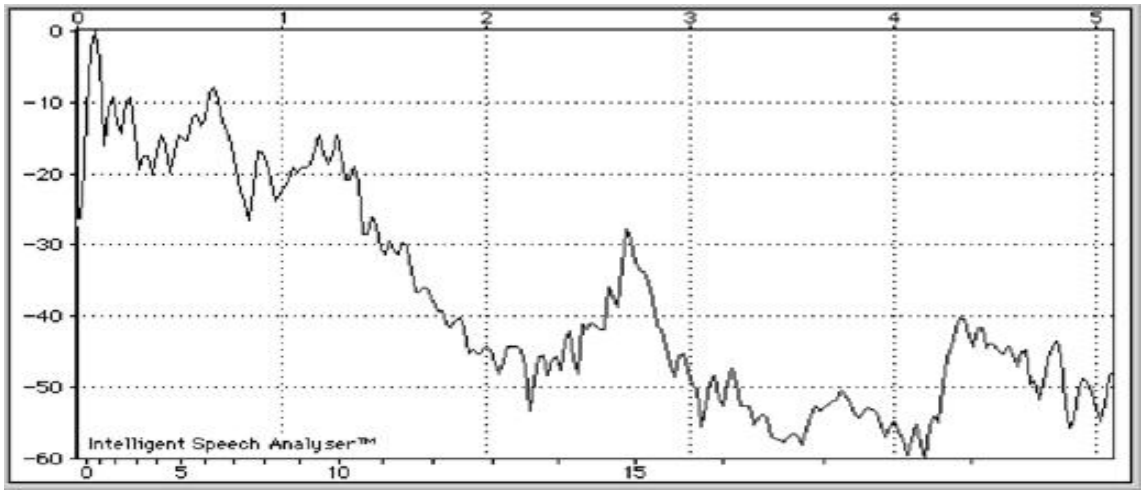


Figure 3b.

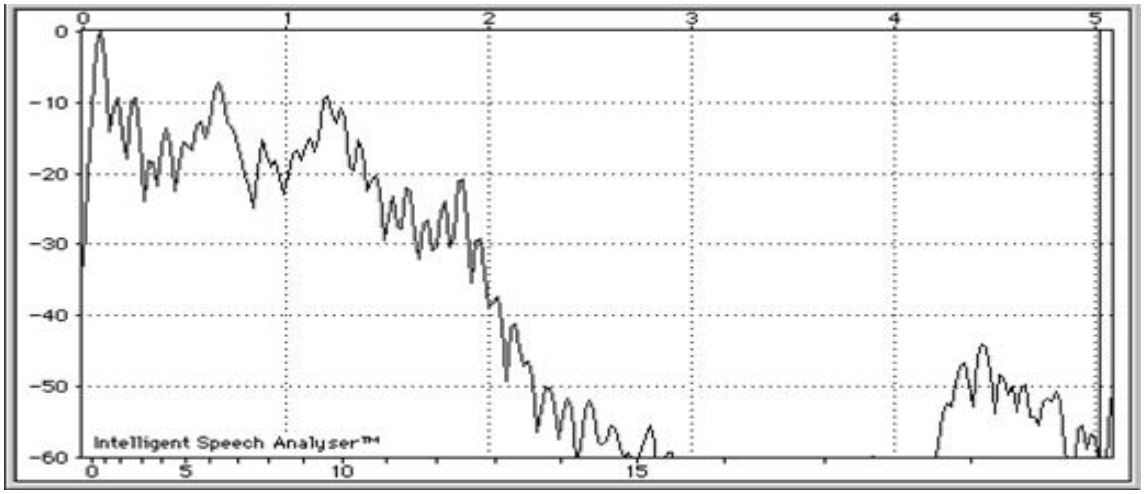


Figure 3c.

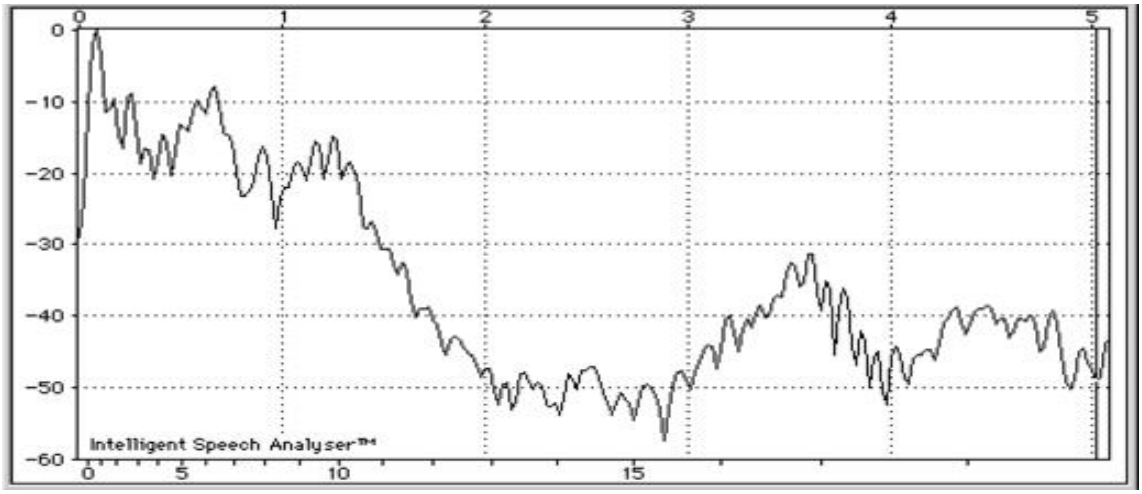
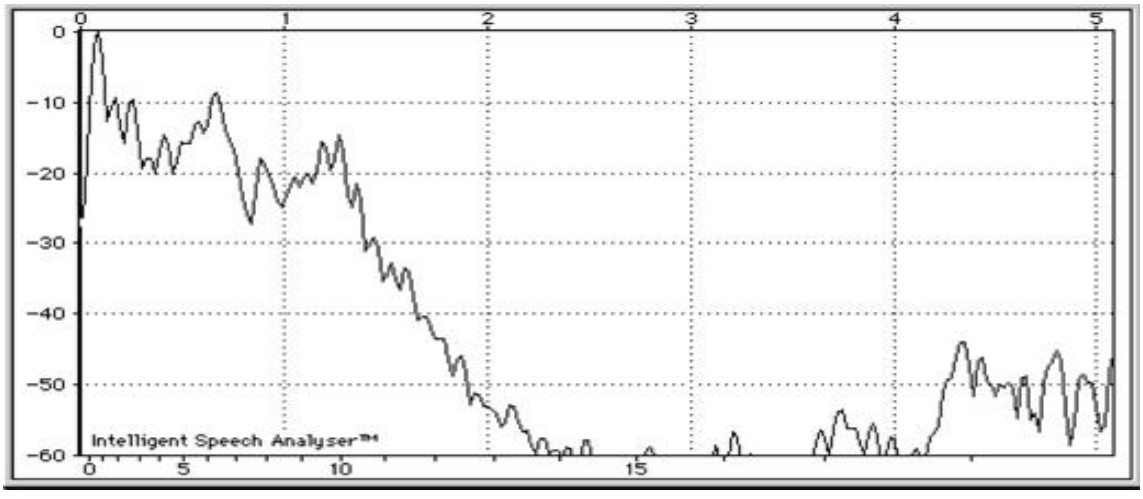


Figure 3d.



3.3. Statistical analyses

Article I. Average values of F0, L_{eq}, duration, formant frequencies F1, F2, F3 and F4, NAQ and alpha ratio of the samples were calculated. Bivariate Pearson correlation coefficients were used to investigate the relations between the acoustic variables. The acoustic variable relations to the valence and psycho-physiological activity level perceived were studied with Multinomial Logistic Regression Analysis. Valence was given values by the researcher as follows: -1 = negative, 0 = neutral, +1 = positive valence, and activity was marked as -1 = low, 0 = medium, +1 = high psycho-physiological activity level. Statistical analyses were computed using SPSS-15 software (SPSS Inc., Chicago, IL).

Article II. Confusion matrices were created for the intended and perceived emotions in the human evaluation test and in the automatic emotion classification experiment. These results were compared with each other in percentages.

Article III. Wilcoxon Signed Rank t-test was used to study the differences between the perception of the original and modified emotional vowel samples. Pearson correlation was carried out to study the relations between valence perceived and Hz value of F3. Intra- and interrater reliability obtained in the listening test was investigated with Cronbach's alpha.

Article IV. The interrelations of acoustic variables in emotional expressions were investigated by Analysis of Covariance (ANCOVA, SPSS-15, Chicago, IL). Valence and psychophysiological activity level were the two main characteristics of expressions studied by assigning them with arbitrary numbers (positive valence and high activity level = 1; a neutral emotional state and medium activity level = 0; negative valence and low activity level = -1). Filter, NAQ, L_{eq} and alpha ratio were set as dependent variables and L_{eq} was set as a covariate in order to study the possible independence of other characteristics of L_{eq}. Bonferroni correction was used in the investigation. Due to problems in the acoustic analysis some of the females' samples were excluded. A confusion

matrix of the emotion samples expressed and perceived was calculated for the listeners' answers. Differences between the genders were studied by Mann-Whitney U test, SPSS-15.

4. Results

4.1. Perception of the samples

Article I. In the listening test with samples of the stress-carrying vowel [a:] samples the agreement level among the listeners concerning the perception of the actual emotional states was 60 % (intrarater reliability was 58 % for the *emotional states* perceived and 61 % for the *valence* perceived). The aim was to study the acoustic characteristics of the samples that had been perceived to reflect certain emotions with an agreement above the chance level. The correct identifications of the intended emotional states in the samples expressed were not of interest. Thus, the results of the listening test were observed according to the responses of the majority of listeners. The percentages of samples perceived as representing certain valences in *males* were: 46.5 % perceived as negative, 40.4 % perceived as neutral and 13.1 % perceived as positive; and in *females*: 48.6 % perceived as negative, 23.6 % perceived as neutral and 27.8 % perceived as positive.

Article II. The results of the listening test of the stress-carrying vowel [a:] samples showed in the confusion matrix that the emotions were discriminated on average with 37.7 % accuracy (**Table 2**). The chance level would have been 20 %.

Table 2. Confusion matrix for the results of the listening test.

	neutral	sadness	joy	anger	tenderness
neutral	**49.67 %**	15.16 %	14.05 %	11.43 %	9.69 %
sadness	21.54 %	**42.91 %**	10.66 %	8.94 %	15.95 %
joy	15.67 %	22.41 %	**28.23 %**	24.48 %	9.21 %
anger	24.49 %	21.77 %	10.44 %	**38.17 %**	5.13 %
tenderness	19.25 %	35.32 %	12.78 %	3.90 %	**28.75 %**

Article III. Cronbach's alphas for interrater reliability (0.586) and for intrarater reliability (0.560) of the listening test were low. The modifications of F3 frequencies in the [a:] vowels did

not correlate with the responses. All modifications of the sample representing anger were perceived as positive in valence and modifications of the sample representing sadness as negative. However, perception of sadness with raised F3 was somewhat (non-significantly) more positive than perception of the other sadness samples. Tenderness was perceived as positive, except that the modification with F3 lowered by 30 % was perceived as negative.

Article IV. The percentage for the recognition accuracy of the mono-pitched vowels [a:], [i:] and [u:] was 50 and for the intrarater reliability 59 in the listening test conducted. The most often correctly recognized emotion was anger, with 68 % accuracy. The poorest recognition percentage was obtained for joy, 37 %.

Joy was chosen for an answer most seldom, with 15 % of the given responses while sadness was chosen for an answer most often with 25 % of all the responses given. Furthermore, there were differences in perceptions between the vowels (**Table 3**): [a:] was conveyed tenderness best, vowels [i:] and [u:] sadness, and all the vowels conveyed anger particularly well. The positive emotions were conveyed poorly especially by vowel [u:].

Table 3. Percentages for correctly recognized emotions expressed in different vowels in the listening test.

Emotion	**[a:]**	**[i:]**	**[u:]**
Neutral	46 %	39 %	40 %
Sadness	42 %	57 %	58 %
Joy	41 %	43 %	28 %
Anger	73 %	64 %	67 %
Tenderness	61 %	48 %	35 %

Understandably, it was easier to recognize valence than the actual emotions from the simple prolonged monopitched vowel samples. Valence was perceived with 70.5 % accuracy. Psychophysiological activity level was perceived with even better accuracy, 76.5 %. Valences conveyed by vowel [a:] and [i:] were recognized somewhat better (76 % accuracy) than those conveyed by vowel [u:] (60 % accuracy). Vowel [i:] was the best conveyer of psychophysiological activity level (86 % accuracy) while the corresponding percentage was 73 % for vowel [a:] and 71 % for vowel [u:]. There were no gender-related differences in perception of the samples expressed. However, females perceived the emotional samples expressed with 52 % accuracy while males perceived them with 48 % accuracy but this difference was non-significant. The answers from the two female exchange students from the Czech Republic were somewhat more accurate compared to the answers of the Finns, but the differences were not significant.

When the perceptions of the samples were studied separately for males and females, it was observed that males perceived 12 samples significantly differently from females. The majority in both genders had perceived 10 of these samples correctly but there was a wide deviation in the males' answers. Two of the 12 samples were perceived differently from the intended emotion by males. Furthermore, the number of unanswered samples was significantly greater in males (N = 92) than in females (N = 43) calculated from all vowels replayed ($p = 0.007$, Mann-Whitney U test, SPSS-15) (see also Waaramaa and Laukkanen 2008).

4.2. Acoustic and statistical results

Article I. The results of the Pearson correlation of the stress-carrying [a:] vowels showed that in both genders NAQ correlated with F0 (r = -0.484 in males, r = 0.268 in females), with L_{eq} (r = -0.579 in males, r = -0.320 in females), and with alpha ratio (r = -0.510 in males, r = -0.309 in females). F0 correlated with L_{eq}, (r = 0.701 in males, r = 0.424 in females), with alpha ratio (r = 0.490 in males, r = 0.553 in females), and with F1 (r = 0.227 in males, r = 0.564 in females). L_{eq} correlated with F1 (r = 0.354 in males, r = 0.423 in females).

Valence. Valence and psycho-physiological activity level were set as dependent variables in the Multinomial Logistic Regression Analysis. According to the Likelihood Ratio Tests, in both genders the perception of valence seemed to be related to NAQ. Valence was also associated with duration and F4 in males, and in females to F0 and F1. It appeared that in *males*, negative valence differed significantly from neutral valence in F2 (F2 was lowest in neutral valence) and positive valence differed significantly from neutral and negative valences in NAQ and duration. Small NAQ value and short duration in males indicated higher probability of perceiving neutral valence than positive or negative valence. In *females*, too, NAQ value was significantly different (smallest) in neutrality. The results suggest that the voice source (reflected in NAQ) may have a role independent of F0 and L_{eq} in the perception of valence. There also appeared to be a tendency for F3 in positive valence to reach higher Hz values than in negative valence in *males*.

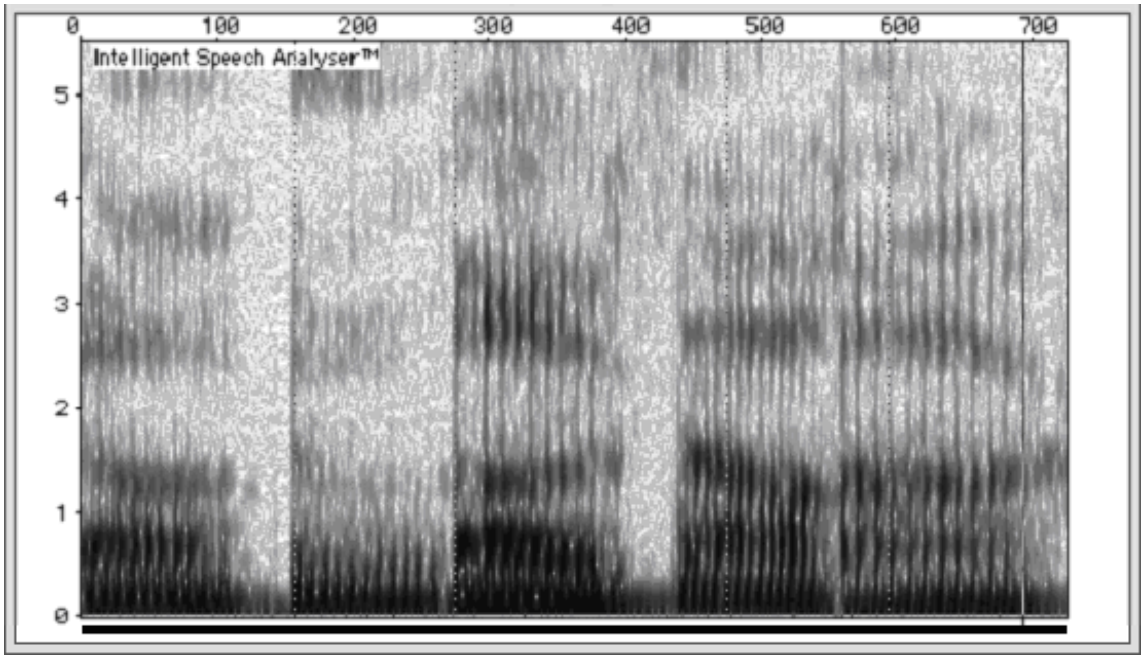
Psychophysiological activity level. The results of the Likelihood Ratio Tests for the perceived psychophysiological activity level showed that perception of activity was mainly associated with L_{eq} and F4 in both genders. Furthermore, F4 seemed to have higher frequency in medium than in high activity level. This result did not appear to be related to the valence of the samples. In *males*, low psychophysiological activity level differed significantly from medium activity level for duration (longer in low activity level) and F2 (higher in low activity level). NAQ was greatest in

low activity level differing significantly in medium and high activity level samples. The results for NAQ were different between genders, most obviously due to the unequal number of samples in different valences. Thus, in *females*, high activity level differed significantly from low and medium activity levels in L_{eq} in the perceived samples.

Figures 4a-b show examples of formant ranges in spectrograms, upper figure from a male and lower figure from a female subject. The samples produced by these subjects were chosen since they were well recognized. Neutrality on the left represents the basic state of the speaker. No attempt is made there to express emotion.

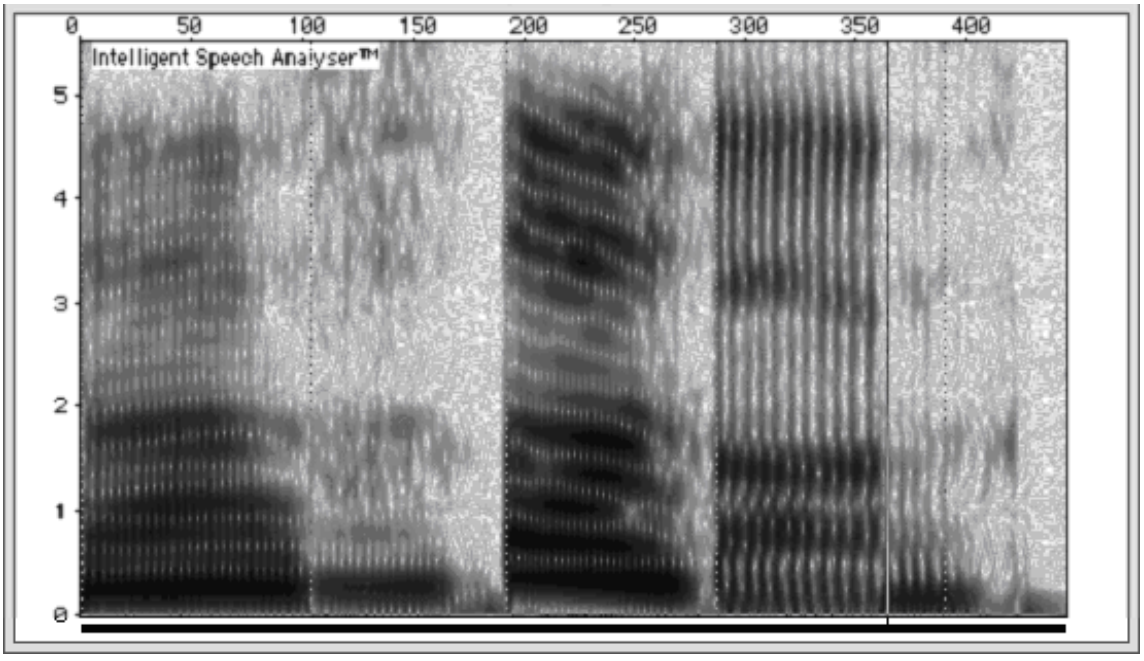
Figures 4a-b. Upper figure: spectrograms of 5 emotions in vowel [a:] produced by a male subject. Lower figure: corresponding spectrograms produced by a female subject. From the left: a neutral emotional state, sadness, joy, anger, and tenderness. Time is on the horizontal axis and kHz on the vertical axis. The spectrograms were made by ISA.

Figure 4a.



neutral sadness joy anger tenderness

Figure 4b.



neutral sadness joy anger tenderness

Article II. The results of the human evaluation test of the stress-carrying [a:] vowels were shown above. The results of the automatic emotion classification experiment are seen in **Table 4**. The results improved somewhat as more high performing features were added to the feature vector. A peak level of 56.9 % (kNN with k = 3) was reached for duration, alpha ratio, S/N ratio, NAQ, and average jitter.

Table 4. Confusion matrix for the results of the automatic emotion classification experiment.

	neutral	sadness	joy	anger	tenderness
neutral	**72.22 %**	11.11 %	16.67 %	0.00 %	0.00 %
sadness	43.75 %	**50.00 %**	0.00 %	0.00 %	6.25 %
joy	21.43 %	14.29 %	**57.14 %**	7.14 %	0.00 %
anger	37.50 %	0.00 %	25.00 %	**25.00 %**	12.50 %
tenderness	6.25 %	18.75 %	6.25 %	6.25 %	**62.50 %**

Article III. Statistically significant results were found for the F3-modified samples of tenderness: that with F3 raised by 30 % was perceived more often as positive than the samples with original ($p = 0.063$) or lowered by 30 % ($p = 0.006$) or completely removed F3 ($p = 0.066$).

Article IV. In the vocal tract function (F1, F2, F3 and F4), the effects of gender were seen in all three monopitched vowels [a:], [i:] and [u:] as expected. Valence perceived was also associated with filter effects in vowel [a:] ($F_{1.51} = 6.18$, $p = 0.016$) differentiating significantly between positive and negative emotions (Bonferroni adjusted $p = 0.016$). The effects of the filter were similar in both genders regarding to valence being somewhat higher in frequency in positive emotions than in negative emotions. As expected, formant patterns differed between vowels in emotional expressions: unlike for [a:], vocal tract function was not significantly related to valence in vowels [i:] and [u:]. Effects of voice source were seen as expected in L_{eq}, which was associated with activity levels in all three vowels, low and medium activity levels: Bonferroni test, $p = 0.001$, and

low and high activity levels: p = < 0.001. In both genders, L_{eq} was highest in anger and lowest in tenderness and sadness. NAQ was significantly related to the psychophysiological activity level ($F_{1,57}$ = 27.9, p = < 0.001) in vowel [a:] also when L_{eq} was set as a covariate. This result suggests that phonation type may vary independently of L_{eq} or that voice source variation may reflect phonation type variation which is not directly related to L_{eq}.

5. Discussion

Perception. The aim of this thesis was to study the role of voice quality in the expression and perception of emotions. The study was carried out by eliminating the effects of those prosodic characteristics which have been shown to be the most obvious conveyers of emotional information, such as F0 variation within the samples. It was presumed that by using vowel samples short enough the desired perceptual "first reaction" would have been caught. The sensitivity of the human ear to temporal changes (Fujimura et al. 1995) would not confuse the perception, either. Thus, it was considered that the sample duration used in the present study was long enough. According to the findings represented in the earlier studies, recognition of emotional information takes place within the first 30-160 ms of the expression, valence being identified faster than the actual emotions (Bostanov and Kotchoubey 2004; Damasio 2003; Izard 2007; Zei Pollermann 2002). This qualification tends to be based primarily on voice quality, which was shown by Bostanov and Kotchoubey (2004) in their study of event-related brain potentials (ERP) of one syllable length utterances. Izard (2007) has stated: "the percept needs to register only in phenomenal consciousness for the basic emotion to become functional" since "a conceptual act is not necessary to enable neural processes". Izard (2007) based this understanding on the results of the neural processing research according to which emotional information is processed within 30 ms, 100 ms before controlled conceptual processing (see also Wambacq 2004). Additionally, Pourtois et al. (2005) have suggested that emotional stimuli, either auditory or visual, recruit the same cortical network. They continue that "audio-visual perception of emotion proceeds covertly in the sense that it does not depend on explicit recognition of the emotions" (Pourtois et al. 2005). As an emotional state is always present in a human mind interacting with new information perceived and cognition processes, it also affects effective incoming stimuli which elicit the target emotions but which will, however, be followed by other emotions within only a few seconds (Izard 2007). For the present

dissertation, this knowledge made it reasonable (for the fourth article) to ask the actors to produce only simple vowel samples to express emotional states.

In earlier studies emotional vocal samples have been recognized with approximately 60 % accuracy (Scherer 1995; 2003). Emotions expressed by visual cues from facial expressions have been recognized with somewhat higher accuracy (Scherer 2003; Ekman 2004; Abelin 2008). Samples with combined acoustic and linguistic information (Lee et al. 2002b) or those with longer duration would naturally have included more indicators of the emotional states expressed than only a single vowel. However, not even a full intonation pattern can always convey a specified emotive meaning in a reliable way; it may rather serve as an indicator of the activation dimension of an emotion (Pakosz 1982; 1983). In the present dissertation, the participants in the listening test agreed with each other and were consistent in the repeated evaluation clearly above chance level. This was the case although the samples used were not selected from the material. Instead, all the samples available were used. In automatic identification, the results reached even a better accuracy level than in a human listening test. According to ten Bosch (2003) the automatic speaker-independent classification performance for three basic emotions (joy, anger, sadness) can reach at its best a 70 % accuracy level the speech units being longer than one phoneme length. Thus, the samples used here appeared to be adequate in duration.

Although the use of short samples was justified, it must be considered that the task was not very easy for the listeners. The recognition of the emotions or valences from the phoneme-length samples was demanding and hence, it could reduce motivation or cause a certain attitude or avoidance action in response to the stimuli. Izard has argued that, for instance, an anger approach may sometimes be considered constructive behavior (Izard 2007). Hence, the interpretation of emotional expression and perception must not be simplified. Izard (2007) has stated: “Although there may be some general characteristics of basic emotions and emotion schemas that justify grouping them by valence or direction of motivated action, broad descriptive terms such as positive

and negative emotions (…) are arbitrary." It is obvious that listeners' motivation or a lack of it is crucial in an evaluation test. According to Izard (2007) the emotion of *interest* is the principle force in organizing consciousness. Similarly, Damasio (2007) has suggested that *enthusiasm* and *discouragement* are always present as background emotions. The importance of the background emotions and their impact on perception of emotional stimuli is obvious since "there is no such thing as affectless mind; affect or emotion is always present", and "the activation of a new emotion involves nonlinear interaction between ongoing emotion and cognition" (Izard 2007; see also Damasio 2007). Against this background it seems inevitable that there may be extensive inter-individual differences in the ability to express and perceive emotions and moreover, this variation may be dependent on their respective intra-individual situations. This aspect has also been pointed out by Abelin (2003).

Sample quality. Emotional samples produced by actors have been claimed to be stereotypical and controlled. Some doubts have been voiced as to whether "acted" samples could not be used in emotion research at all. This acquisition raises another question about how genuine our habitual emotional states ever are if they are mixed with other ongoing emotions quite randomly. Does a pure emotion exist at all and if it does, what is its manifestation like? At least in the present dissertation, even though the emotional states were acted, they were recognizable by the listeners. Hence, there have to be some cues, either universal or cultural, which the listeners believed were expressing certain emotional states. Furthermore, how would acting be possible if emotions could not be imitated and presented recognizably on the stage (see also Scherer 2003)? According to Ohala (1996) emotional signals may be viewed as designed to influence receiver's behaviour in order to benefit the signaller. Thus, emotional expressions are not only passive reflections of the psychophysiological state of the speaker but also actively used tools to produce favourable responses (Ohala 1996). There are also several institutions e.g. authorities, other social systems and the social environment by which our social (emotional) behavior is controlled (Banse and Scherer

1996; Feldman Barret 2006). Those of us who do not evoke any contradictory emotions in other people are said to have good social competence or good social skills. To have competence or skills requires control over the subject. Thus, it does not seem to be reasonable to claim that in social life emotions are uncontrolled and hence, "real". Thus, there was no reason not to use acted emotional samples like the materials in the current dissertation since the samples used appeared to be adequate in quality.

Emotions. Neutrality was used especially to scale the differences in the parameters regarding both ends, positive and negative, of the scale. However, it cannot be assumed that neutrality does not include any emotion at all, but it may be considered that in neutrality no special emotional cue is emphasized. This may be also the explanation why neutrality was confused with negative emotions: if there are no affective cues in the speech at all it would sound without any doubt mostly unfriendly. Neutral utterance has been shown to be quite monotonous in contour and shallow in range (Morrison 2007).

Tenderness can be considered similar to basic emotions since it is very close to the nursing instinct, which is inherent both in humans and animals. Sadness was often expressed or perceived in samples which unexpectedly high formant frequencies. Naturally, sadness expressed in this way does not sound depressive; it sounds frustrated or reminiscent of grief. Frick (1986) has suggested that frustration is one of the forms of anger. The other form would be threat. According to Morrison et al. (2007) frustration has similar but somewhat smaller physiological effects to anger; F0 is higher than in neutral emotional speech.

Joy appeared to be most difficult to recognize for the listeners. Kotlyar and Morazov (1976) (see also Sundberg 1987) and Lee et al. (2004) have reported similar results. It may be that from the survival-evolutionary point of view it has been unnecessary to learn to recognize joy or happiness very quickly and hence, dissipate extra energy on reflecting stimuli which were not life threatening. However, it may have been important to react quickly to highly aroused stimuli. Perhaps this is the

reason why anger and joy were often confused with each other, joy being often perceived as anger. The confusion occured most obviously due to their similar activity level, which is typically high in both. This assumption is supported by Pakosz (1982; 1983). According to Pakosz, similarly loaded activation dimensions between two emotional expressions lead to their misidentification by confusing them in perception. In the first study of the present dissertation it was notable, that of the samples produced by males only, 13 % were perceived as positive while those produced by females almost 28 % were perceived as positive. Similar results were obtained by Bonebright et al. (1996) in their study of gender stereotypes in the expression and perception of vocal affects. They found that samples of happiness portrayed by females were identified more accurately than those produced by males. Moreover, male anger and fear were better recognized than female anger or fear (Bonebright et al. 1996). In their event-related potentials (ERP) investigation Schirmer and Kotz (2002), asked the participants to judge the valence of the prosody of a German verb as positive, neutral or negative. Similarly in the second block, the participants were asked to judge the emotional meaning of the word. The researchers found an interaction of emotional prosody and word meaning in females but not in males. Instead, males appeared to process the meaning of a word and the emotional prosody independently of each other. They also suggested that females are faster and more accurate in judging emotional information than males and they are also more strongly engaged with social interactions than males (Schirmer and Kotz 2002; Schirmer et al. 2002; see also Besson et al. 2002; Fecteau et al. 2005a; Schirmer and Simpson 2008). Schirmer et al. (2005; see also Schirmer and Simpson 2008) have also noted that these gender differences emerge at an early, automatized stage of information processing. The gender difference may lie in females' somewhat more bilateral brain processing in response to emotional vocalizations (Schirmer and Simpson 2008). A recent study by Gur et al. (2002) concentrated on anatomic gender differences in the brain. They suggested that there may be significant neuroanatomical differences in the regional volumes in the frontal lobes where orbital frontal regions were relatively larger in females than in males.

According to the researchers, this has a decisive effect on the capability to modulate amygdala input and emotional behaviour, particularly aggression. A reduced frontal volume was prone to cause antisocial behaviour and even psychopathy in healthy males (Gur et al. 2002).

Interestingly, in the automatic speaker independent identification, joy was better (57 % vs. 28 % accuracy) and anger was less well recognized than in human listening test (25 % vs. 38 % accuracy) (*Article II*). Naturally, from the evolutionary-survival point of view it is understandable that anger is well recognized by man and recognizing joy may not have been as necessary to survival. Again, the question remains why the automatic identification system could not identify anger as well as humans or as well as joy (*Article II*). Furthermore, the semi-synthetic anger samples (*Article III*) were perceived as positive in valence in all modifications in the human evaluation test. This confusion between joy and anger may be due to their typically high pitch and the large amount of energy in high frequency range areas in both of these emotions (e.g. Murray and Arnott 1993; Chuenwattanapranithi et al. 27.10.2008). However, some kind of mismatch between humans and automatic identification in discriminating joy and anger was evident. One possible explanation for the mismatch may be that the automatic classifiers are not (yet) able to imitate the human brain, which, in turn, is tuned in a special way to the human voice, different from any other sounds (Fecteau et al. 2004a). At the moment this differentiation is conceivable only for the human ear. In automatic classification the quantitative parameters were utilized effectively: only two parameters, duration and alpha ratio, were needed to obtain highly satisfactory results for the recognition of the emotions. By adding jitter, shimmer and S/N ratio to the analysis the results improved only marginally. Taken together, it seems that humans and the automatic identification pick different characteristics from emotional vocal signals for their identification. The automatic classification may be more precise than human listeners, it may not need to get used to the sounds before recognizing them, i.e. scaling of the ear is not needed; nor does it tire as humans do; it is not bothered by any thoughts or emotions as are humans, which may disturb the listening task;

nevertheless, automatic identification does not understand the signals perceived as humans do.

In a cross-linguistic study of the interpretation of emotional prosody, Abelin and Allwood (2000) investigated Swedish, English, Finnish and Spanish listeners' responses to emotional expressions uttered by a Swedish-speaking subject. The results showed that Finnish speakers were the poorest at recognizing happy utterances. Anger, however, was recognized by the Finns better than the average. It could be speculated that Finnish culture is relatively permissive in expressing negative emotions. Also, expressions of positiveness are often interpreted as excessively feminine or insinuating, naive or as lack of judgment. There are also sayings prohibiting excessive laughter, or otherwise all will end badly. Gravity and seriousness have been highly valued. Such a cultural background may direct the emotional behavior and character of the population.

Voice quality and valence. Perception of valence from *females*' samples appeared to be associated with F0 and F1 (in the first article). F0 was higher in positive than in negative valence in the samples with high effort level and lower in the positive samples with low effort level. F1 was higher in positive valence compared to negative valence, which may imply shortening of the vocal tract by smiling. The role of formant frequencies in valence perception did not tend to be as important as was hypothesized. No statistically significant emotion patterns for formant use were found. Their inter-individual differences appeared to vary over a wide range implying a wide variety of means and their relations through which valence was expressed in different activity levels. It should also be kept in mind that formants have an important linguistic task which may decrease their availability as carriers of emotional information. Hence, it is apparent that valence coding depends on multiple parameters (see also Mozziconazzi 1998; Aubergé and Cathiard 2003). If a parameter does not have statistical significance there may be such reasons as: 1) the parameter does not affect the object researched; 2) the parameter may be used but not necessarily always if there are also other means to express the same thing, e.g. emotion; 3) the parameter may be used in a certain degree of expressed emotion, e.g. when expressing hot anger but not cold anger, and the

inter-individual range of the scales of expressions may vary; and 4) the parameter may be normally used in the expression of a certain emotion, however, the emotion samples may not always be successful in their expressiveness or the sample may not be perfect just at the particular point at which is studied.

However, there appeared to be a tendency to perceive positive valence from samples with somewhat higher F3 frequency. The modifications of F3 in the third article showed that although voice quality was hypofunctional in tenderness and the amount of energy was relatively low in the higher frequency range area in the spectrum, the samples expressing tenderness with raised F3 were nevertheless perceived more often as positive than were the other modifications. Yet, the amount of energy in the higher frequency range was greater in tenderness than in sadness. This may have perceptual importance in valence coding. Instead, if the amount of energy is low in the higher frequency range it may not affect on valence perception as much since it may not be sufficiently audible.

Different vowels seemed to be more or less successful in conveying emotional information. The open back vowel [a:] conveyed tenderness well, the front vowel [i:] joy and also sadness, and vowel [u:] sadness. The positive emotions tenderness and joy were carried least effectively by vowel [u:] in both genders. Again, anger was remarkably well recognized in all vowels studied, implying its importance that it is not vowel dependent and may be more related to voice source characteristics, not vocal tract effect. Instead, most likely the other emotions were dependent on resonances which differ between the vowels. In [a:] the formant structure is evenly spread and in [i:] F2, F3 and F4 are all relatively high in frequency. The formant structure may give vowel [a:] more freedom to vary and that of [i:] gives it a bright timbre when expressing positive valence and a darker timbre perhaps by lowering the formant amplitudes in the spectrum (for instance by either making the voice source more hypofunctional or by increasing the frequency distance between F2 and F3). This difference between the vowels is used to advantage in songwriting by using those

vowels in the text which support the spirit or the mood of the music. The desired timbre may be achieved e.g. by using a lot of /a/ and /e/ vowels for happy songs and e.g. /u/ and /y/ for darker coloured messages (see e.g. Pierrehumbert 1991). According to Ohala's idea of sound symbolism there are cross-language patterns in tone, vowel and consonant use in the sound-symbolic vocabulary (Ohala 1984, 1997). The result for [u:] may hence suggest that vowel [u:] *per se* may be interpreted as a carrier of darkness. Moreover, in a rounded vowel like /u:/ it is difficult to fulfil an articulatory setting in which the lips are retracted as in smiling while expressing positive emotions. In the pronunciation of /u:/ the lips are more or less protruded. Thus, production of /u/ seems in a way to be the opposite of smiling. This may also be one reason for the impression or sound symbolism which we tend to glean from vowel [u:].

In the fourth study (*Article IV)* the number of unanswered samples was significantly greater in males than in females ($p = 0.007$). As Schirmer and Kotz (2002) have suggested that females are faster in emotional judgments than males, this result could by interpreted such that there may be more hesitation in this decision-making on emotional evaluation in males than in females.

As the reported studies appear to suggest NAQ may vary independently of F0 and L_{eq}. This would suggest that phonation type may be used independently along the axis hypofunctional – hyperfunctional (perceived as breathy - pressed voice quality). This is reflected in NAQ from high to low values (Alku et al. 2002; Cowie and Cornelius 2003). In the first study gender differences were obtained from NAQ as well as in the studies by Airas and Alku (2006) and Waaramaa et al. (2008, in press). However, this was not the case with the material investigated in the fourth study, where no gender differences were found for NAQ in the mono-pitched expressions. This result may reflect differences in the F0 control systems between the genders, not differences in the voice quality. This may imply that the phonation type changes in accordance with F0. The statistical results seemed further to suggest that valence may be mainly carried by NAQ in both genders. An earlier study by Waaramaa et al. (2007) concluded similarly that NAQ may have an effect on the

perception of valence. However, the results of NAQ as a carrier of valence have to be interpreted with caution since the activity levels expressed most obviously confuse these results (high activity level in positively valenced joy and negatively valenced anger and low activity level in positively valenced tenderness and negatively valenced sadness). It seems that NAQ may not carry valence at all. Additionally, an unequal number of different emotion samples in the first study may have an effect on the results of NAQ. The unequal number between the emotions was inevitable since it was not predetermined which emotional states the listeners should hear; instead the listeners chose the emotions they really heard, and hence, the resulting numbers were not equal.

The fourth study concerned different vowels, and, as expected, the alpha ratio was found to be significantly higher in vowel [a:] in all emotions studied than in the other vowels [i:] and [u:]. The higher alpha ratio in [a:] may have some importance for automatic identification accuracy. In the present dissertation only vowel [a:] was studied in the automatic classification. It would be reasonable also to investigate the other vowels in order to see whether the automatic classification would be as accurate in them as it was in the identification of emotions from vowel [a:]. Lee et al. (2004, see also Lee et al. 2002a) studied emotional states of anger, happiness, sadness and neutrality using HMM (Hidden Markov Models) classifiers and 5 phoneme classes, vowel, glide, nasal, stop and fricative sounds. The authors hypothesized that different emotional states distinctly affect different phonemes. They found that vowel sounds were good indicators for emotion recognition with 72 % accuracy, and that vowels with open vocal tract like "/AA/" appeared to have more scope for the emotional coloring of expressions than vowels with other kinds of characteristics. (See the discussion above about the formant structure of vowel [a:] and its greater freedom to vary compared to the other vowels studied in the present dissertation.)

Although it was difficult to find any certain acoustic pattern for different vocal emotional expressions it is important to understand that emotive communication forms are not arbitrary. "The arbitrary signs are necessarily conventional but the conventional signs are not necessarily arbitrary."

(Fónagy and Magdics 1963, 298; the quotation was originally published in “Über die Eigenart des sprachlichen Zeichens”, Lingua 1956: 6, 67-88.)

6. Conclusions

1. The differences between the terms affect and emotion have been defined and a new concept of the emosphere proposed.

2. It appeared to be possible to identify emotional valence from vowel samples as short as ~150 ms in duration and the actual emotions from vowel samples ~2400 ms in duration.

3. The automatic classification of emotional phoneme length stimuli has also been shown to be possible with a good accuracy rate. Human listeners' accuracy in recognizing emotional content in speech was clearly below that of the machine identification in an experimental situation. However, anger was better recognized by humans and joy by the automatic classification. Further, it seems that humans and automatic evaluation use different parameters in emotion recognition.

4. Voice source did not only reflect variations of F0 and L_{eq} but appeared to have an independent role in expression, reflecting phonation types measured in NAQ values.

5. Formant frequencies F1, F2, F3 and F4 were related to valence in vowel [a:]. The perception of positive valence tended to be associated with higher frequency of F3 but no clear pattern could be detected, probably reflecting the differences in formant use on different activity levels.

6. Mono-pitched vowels [a:], [i:] and [u:] differed in their capacity to carry emotional information. In both genders, vocal tract effects were associated with valence in vowel [a:], being somewhat higher in frequency in positive valence than in negative valence. Significance was found only for L_{eq} in vowels [i:] and [u:] in emotional expressions. This may be due to different use of voice source and filter characteristics in different vowels or due to the fact that the same phonatory or articulatory characteristics have different acoustic consequences in the vocal tract setting in different vowels. However, the results for vowel [u:] may be biased since the number of [u:] samples in females was smaller than the vowel data of [a:] and [i:].

7. In both genders, psycho-physiological activity level was coded mainly through L_{eq}.

8. Perception of valence tends to be a complex multilevel parameter with wide individual variations (i.e. due to differences in the individual emosphere).

9. The perceptual effects of the interplay between voice source and formant frequencies in different vowels warrant further study by modified synthetic samples yet preserving natural sound.

10. There may be more hesitation in males than females in making decisions on the quality of emotional information perceived. Whether the reason for this is simply motivational or due to gender differences in brain processing warrants further study.

7. Acknowledgements

First of all I want to express my warmest thanks to Professor Anne-Maria Laukkanen (Department of Speech Communication and Voice Research, University of Tampere) and Professor Paavo Alku (Department of Signal Processing and Acoustics, Helsinki University of Technology). It has been a great honor to have these prominent scientists as supervisors in my dissertation. Their outstanding knowledge and teaching has been an indispensable gift they have given to this dissertation. Working in the same department with Professor Anne-Maria Laukkanen has provided me with a special opportunity to benefit from her patient guidance. There was no question to which she did not make time to give a thorough answer.

I am also the most grateful to Professor Olli Aaltonen (University of Helsinki) and Professor Krzysztof Izdebski (Stanford University, California) who have been the pre-examiners for my work. These specialists have given me valuable advice on the present dissertation.

I wish to thank all the co-authors I have had the opportunity the work with, Matti Airas D.Sc.(Tech.) (Helsinki University of Technology), Professor Tapio Seppänen, Juhani Toivanen Ph.D. and Eero Väyrynen M.Sc. (MediaTeam, University of Oulu) and special laboratory technician Jarmo Helin (University of Tampere) for his collaboration with the materials for this dissertation. Special thanks to Jarkko Niemi Ph.D. and all the participants who made the listening tests possible. Without the help of the specialists in statistics Hanna-Mari Puuska (née Pasanen) M.A. and Jyrki Ollikainen M.A. this work would have been much harder to carry out. Great thanks to both of them. Virginia Mattila M.A. checked and corrected the English language in all the contributions related to this dissertation. I wish to thank her for the highly professional work she has done.

I am deeply grateful to the whole staff of the Department of Speech Communication and Voice Research for sharing their knowledge when ever I have needed it. Their friendly attitude has been of great value. Especially the discussions with Irma Ilomäki Ph.D., Professor Anna-Maija

Korpijaakko-Huuhka, Tarja Kukkonen Ph.L., M.A.(Ed.), M.Sc., Elinita Mäki M.Sc. and Leena Rantala Ph.D. have helped me with the present work.

Last but not least I want to thank my dear husband Heikki Mäki-Kulmala Ph.D. who has been the most supportive in my efforts to understand and study the phenomenon this work has focused on. His comments have helped me to see the wholeness from its parts and his great learning has opened me many locks on my path of learning. I hope this work may also guide the paths of my dear children, Veera and Joonas, to make wise choices and encourage them in their future lives.

8. Financial Support

This research project was financially supported by Academy of Finland (project numbers 200859, 200807 and 200997) and The University of Tampere.

9. References

Aaltonen O. The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. Journal of Phonetics 1985; 13, 1-9.

Aaltonen O. Vowel perception: behavioural and psychophysiological experiments. Doctoral dissertation. Centre of cognitive neuroscience, and department of Finnish and general linguistics (Phonetics laboratory). University of Turku. Turku 1997.

Abelin Å. Anger or fear? Cross-cultural multimodal interpretations of emotional expressions. In: K. Izdebski (Ed.) Emotions in the Human Voice. Vol. I. Plural Publishing. San Diego. 2008, 65-73.

Abelin Å. Interpretation of emotions in natural speech – a comparison between written, auditive and gestural information. 15th ICPhS, Barcelona 2003.

Abelin Å, Allwood J. Cross linguistic interpretation of emotional prosody. ISCA ITRW Workshop on Speech and Emotion. Newcastle, Northern Ireland, United Kingdom. September 5-7, 2000. (www.ling.gu.se/~abelin/abelin.pdf)

Addington DW. The relationship of the selected vocal characteristics to personality perception. Speech Monographs 1968: XXXV, 492-503.

Airas M, Alku P. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient. *Phonetica.* 2006; 63: 26-46.

Alku P. An automatic inverse filtering method for the analysis of glottal waveforms. Academic dissertation. Helsinki University of Technology. Faculty of Electrical Engineering. Acoustics Laboratory. Otaniemi. 1992.

Alku P, Airas M, Björkner E, Sundberg J. An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. Journal of The Acoustical Society of America 2006: 120: 2, 1052-1062.

Alku P, Bäckström T, Vilkman,E. Normalized amplitude quotient for parametrization of the glottal flow. Journal of The Acoustical Society of America 2002: 112: 2, 701-710.

Alku P, Vilkman E, Laukkanen A-M. Estimation of amplitude features of the glottal flow by inverse filtering speech pressure signals. Speech Commununication. 1998a: 24, 123-132.

Alku P, Vilkman E, Laukkanen A-M. Parameterization of the voice source by combining spectral decay and amplitude features of the glottal flow. Journal of Speech, Language, and Hearing Research. 1998b: 41: 5, 990-1002.

Alku P, Vilkman E, Laine UK. Analysis of glottal waveform in different phonation types using the new IAIF-method. In: Proceedings of the XII International Congress of Phonetic Sciences (ICPhS'91). Aix-en-Provence, France, August 19-24, 1991: 4, 362-365.

Aubergé V, Cathiard M. Can we hear prosody of smile? Speech communication 2003: 40:1-2, 87-97.

Banse R, Scherer KR. Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology 1996: 70: 3, 614-636.

Belin P, Feacteau S, Bédard C. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Science 2004: 8:3, 129-135.

Besson M, Magne C, Schön D. Emotional prodsody: sex differences in sensitivity to speech melody. Trends in Cognitive Sciences 2002: 6: 10, 405-407.

Birkett PB, Hunter MD, Parks RW, Farrow TF, Lowe H, Wilkinson LD, Woodruff PW. Voice familiarity engages auditory cortex. Neuroreport 2007: 18: 13, 1375-1378.

Blood GW, Mahan BW, Hyman M. Judging personality and appearance from voice disorders. Journal of Communication Disorders 1979: 12, 63-68.

Bonebright TL, Thompson JL, Leger DW. Gender stereotypes in the expression and perception of vocal affect. Sex roles. 1996:34: 5-6, 429-445.

ten Bosch L. Emotions, speech and the ASR framework. Speech Communication 2003: 40: 1-2, 213-225.

Bostanov V, Kotchoubey B. Recognition of affective prosody: Continuous wavelet measures of event-related brain potentials to emotional exclamations. Psychophysiology. 2004; 41: 259-268.

Campbell N, Mokhtari P. Voice quality: the 4th prosodic dimension. Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona, 3-9 August 2003, 2417-2420.

Chuenwattanapranithi S, Xu Y, Thipakorn B, Maneewongvatana S. Encoding emotions in speech with the size code — A perceptual investigation. http://www.phon.ucl.ac.uk/home/yi/publications.html. 27.10.2008.

Chusid JG. Correlative neuroanatomy & functional neurology. 16th ed. Lange Medical Publications. Los Altos, California. 1976.

Cleveland T. The acoustic properties of voice timbre types and the importance of these properties in the determination of voice classification in male singers. STL-QPSR. 1976,: 17: 1, 17-29.

Cowie R, Cornelius RR. Describing the emotional states that are expressed in speech. Speech Communication 2003; 40; 1-2: 1-33.

Cummings KE, Clementes MA. Analysis of the glottal excitation of emotionally styled and stressed speech. Journal of The Acoustical Society of America. 1995: 98: 1, 88-98.

Damasio A. Brain and mind: from medicine to society. Conference lecture (video) in "Brain and mind: from medicine to society", Barcelona, Spain. 24th of May 2007. 1/2http://www.youtube.com/watch?v=KbacW1HVZVk&NR=1. 9.5.2008.

Damasio A. Looking for Spinoza: Joy, sorrow, and the feeling brain. A Harvest Book, Harcourt, Inc. USA. 2003.

Damasio AR. The feeling of what happens: Body and emotion in the making of consciousness. Harcourt Brace. New York. 1999.

Darwin C. The expression of the emotions in man and animals. The Thinker's Library No. 47. Watts & Co. Limited. Great Britain. 1934 (1872).

Ekman P. Emotions Revealed Recognizing faces and feelings to improve communication and emotional life. Owl Books. New York. 2004.

Fant G. Acoustic theory of speech production. With calculations based on X-ray studies of Russian articulations. (2nd ed.). The Hague: Mouton. 1970.

Fant G, Lin Q. Glottal source-vocal tract acoustic interaction. STL-QPSR. 1987: 1, 13-27.

Fecteau S, Armony JL, Yves J, Belin P. Is voice processing species-specific in human auditory cortex? An fMRI study. NeuroImage 2004a: 23, 840-848.

Fecteau S, Armony JL, Yves J, Belin P. Judgment of emotional nonlinguistic vocalization: Age-related differences. Applied Neurophysiology 2005a: 12: 1, 40-48.

Fecteau S, Armony JL, Yves J, Belin P. Priming of non-speech vocalizations in male adults: The influence of the speaker's gender. Brain and cognition 2004b: 55, 300-302.

Fecteau S, Armony JL, Yves J, Belin P. Sensitivity to voice in human prefrontal cortex. Journal of Neurophysiology 2005b: 94, 2251-2254.

Feldman Barrett L. Are emotions natural kind? Perspectives on Psychological Science. 2006: 1:1.

Fitch WT. The evolution of speech: a comparative review. Trends in Cognitive Sciences. 2000: 4: 7, 258-267.

Fónagy I, Magdics K. Emotional patterns in intonation and music. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 1963: 16, 293-326.

Frøkjær-Jensen B, Prytz S. Registration of voice quality. Brüel & Kjær Technical Review 1973: 3: 3-17.

Fujimura O, Cimino A, Sawada M. Voice quality control within sentence: Expressive effects of source spectral envelope change. In: Vocal fold physiology. Voice quality control. Eds. Fujimura O. and Hirano M. Singular Publishing Group, Inc. San Diego, California. 1995.

Frick RW. The prosodic expression of anger: differentiating threat and frustration. Aggressive Behavior 1986: 12, 121-128.

Gauffin J, Sundberg J. Spectral correlates of glottal voice source waveform characteristics. Journal of Speech and Hearing Research 1989: 32, 556-565.

Gobl C, Ní Chasaide A. The role of voice quality in communicating emotion, mood and attitude. Speech Communication 2003: 40: 1-2, 189-212.

Granström B, Nord L.Ways of exploring speaker characteristics and speaking styles. Proceedings of the 12th International Congress of Phonetic Sciences, 19-24 August Aix-en-Provence 1991, 278-281.

Guerrero LK, Andersen PA, Trost MR. Communication and emotion: Basic concepts and approaches. In: Andersen PA, Guerrero LK, editors. Handbook of communication and emotion. Research, theory, applications and contexts. USA: Academic Press. 1998, 3-27.

Gur RC, Gunning-Dixon F, Bilker WB, Gur RE. Sex differences in temporo-limbic and frontal brain volumes of healthy adults. Cerebral Cortex. 2002: 12: 9, 998-1003.

Holmberg EB, Hillman RE, Perkell JS. Glottal airflow and transglottal air pressure measurements for male and female speakers in low, normal, and high pitch. Journal of Voice. 1989: 3: 4, 294-305.

Holmberg EB, Hillman RE, Perkell JS. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. Journal of The Acoustical Society of America. 1988; 84: 2, 511-529.

Howard DM, Tyrell AM. Psychoacoustically informed spectrography and timbre. Organised Sound 1997: 2: 2, 65 – 76.

Imberty M. The question of innate competencies in musical communication. In: The origins of music. Eds. Wallin NL, Merker B and Brown S. Massachusetts Institute of Technology. USA. 2001, 449-462.

Izard CE. Basic emotions, natural kinds, emotion schemas, and a new paradigm. Perspectives on Psychological Science 2007: 2: 3, 260-280.

Kohler E, Keyers C, Umiltá MA, Fogassi L, Gallese V, Rizzolatti G. Hearing sounds, understanding actions: Action representation in mirror neurons. Science 2002: 297, 846-848.

Kotlyar GM, Morozov VP. Acoustical correlates of the emotional content of vocalized speech. Soviet Psychology and Acoustics 1976: 22, 208-211.

Ladd DR, Silverman KEA, Tolkmitt F, Bergmann G, Scherer KR. Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. Journal of The Acoustical Society of America 1985: 78: 2, 435-444.

Laukkanen A-M. On speaking voice exercises. A study on the acoustic and physiological effects of speaking voice exercises applying manipulation of the acoustic-aerodynamic state of the supraglottic space and artificially modified auditory feedback. Doctoral dissertation. Medical School. University of Tampere, Finland. 1995.

Laukkanen A-M, Alku P, Airas M, Waaramaa T. The role of voice in the expression and perception os emotions. In: K. Izdebski (Ed.) Emotions in the Human Voice. Vol. I. Plural Publishing. San Diego. 2008, 171-184.

Laukkanen A-M, Vilkman E, Alku P, Oksanen H. On the perception of emotions in speech: the role of voice quality. Scandinavian Journal of Logopedics, Phoniatrics, Vocology 1997: 22: 4, 157-168.

Laukkanen A-M, Vilkman E, Alku P, Oksanen H. Physical variations related to stress and emotional state: a preliminary study. Journal of Phonetics 1996: 24, 313-335.

Laver J. The phonetic description of voice quality. Cambridge: Cambridge University Press. 1980.

Lee CM, Narayanan S, Pieraccini R. Classifying emotions in human-machine spoken dialogs. In ICME. Lausanne, Switzerland. 2002a.

Lee CM, Narayanan S, Pieraccini R. Combining acoustic and language information for emotion recognition. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). Denver, Colorado. 2002b.

Lee CM, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S. Emotion recognition based on phoneme classes. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). Jeju Island, Korea. 2004.

Leinonen L, Hiltunen T, Linnankoski I, Laakso M-L. Expression of emotional-motivational connotations with a one-word utterance. Journal of The Acoustical Society of America 1997: 102: 3, 1853-1863.

Lieberman P. The speech of primates. The Hague: Mouton. 1972.

Lieberman P, Klatt DH, Wilson WA. Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. Science 1969: 164, 1185-1187.

Lieberman P, Michaels SB. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. Journal of The Acoustical Society of America 1962: 34: 7, 922-927.

Lotman J. On the semiospere. Sign System Studies 2005: 33: 1, 207-229.

Luria AR. The working brain. An introduction to neuropsychology. Basic Books. 1973.

MacNeilage PF. The frame/content theory of evolution of speech production. Behavioral and brain sciences. 1998: 21, 499-546.

Makarova V, Petrushin VA. Phonetics of emotion in Russian speech. Proceedings of the 15th International Congress of Phonetic Sciences. 3-9 August Barcelona 2003: 2857-2860.

McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S. Approaching automatic recognition of emotion from voice: a rough benchmark. Proceedings of the ISCA Workshop on Speech and Emotion, Belfast. 2000, 207-212.

Morrison D, Wang R, De Silva LC. Ensemble methods for spoken emotion recognition in call-centres. Speech Communication 2007: 49, 98–112.

Mozziconazzi SJL. Speech variability and emotion: Production and perception. Doctoral dissertation. Netherlands: Technische Universiteit Eindhoven, 1998.

Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of The Acoustical Society of America 1993: 93: 2, 1097-1108.

Nishitani N, Hari R. Viewing lip forms: Cortical dynamics. Neuron 2002: 36: 6, 1211-1220.

Nussbaum MC. The Fragility of Goodness. Luck and ethics in Greek tragedy and philosophy. (2nd ed.). USA: Cambridge University Press. 2001.

Ohala JJ. An ethological perspective on common cross-language utilization of F0 of voice. Phonetica. 1984: 41, 1 - 16.

Ohala JJ. Ethological theory and the expression of emotion in the voice. Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP), Philadelphia. Wilmington: University of Delaware. 3rd-6th October 1996: 3, 1812-1815.

Ohala JJ. Sound symbolism. Proceedings of the 4th Seoul International Conference on Linguistics (SICOL). 11- 15th of August 1997, 98-103.

Pakosz M. Attitudinal judgements in intonation: some evidence for a theory. Journal of Psycholinguistic Research 1983: 12: 3, 311-326.

Pakosz M. Intonation and attitude. Lingua 1982: 56, 153-178.

Pierrehumbert JB. Music and the phonological principle: Remarks from the phoneticians' bench. In: Sundberg J, Nord L, Carlson R. (Eds.) Music, language, speech and brain. Proceedings of an International Symposium at the Wenner-Gren Center. Stockholm, 5-8- September 1990. Macmillan Academic and Professional Ltd. Great Britain 1991, 132-145.

Polivy J. On the induction of emotion in the laboratory: Discrete moods or multiple affect states? Journal of Personality and Social Psychology 1981: 41: 4, 803-817.

Pollack I, Rubenstein H, Horowitz A. Communication of verbal modes of expression. Language and Speech 1960: 3: 3, 121-130.

Pourtois G, de Gelder B, Bol A, Crommelinck M. Perception of facial expressions and voices and of their combination in the human brain. Cortex. 2005: 41, 49-59.

Richman B. How music fixed "nonsense" into significant formulas: on rhythm, repetition and meaning. In: The origins of music. Eds. Wallin NL, Merker B and Brown S. Massachusetts Institute of Technology. USA. 2001, 301-314.

Rizzolatti G, Arbib MA. Language within our grasp. Trends in Neurosciences 1998: 21: 5, 188-194.

Rothenberg M. Acoustic interaction between the glottal source and the vocal tract. In: Vocal fold physiology. (Eds.) Stevens KN, Hirano M. University of Tokyo Press. 1981, 305-328.

Rothenberg M. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. Journal of The Acoustical Society of America. 1973: 53: 6, 1632-1645.

Scherer KR. Expression of emotion in voice and music. Journal of Voice 1995: 9: 3, 235-248.

Scherer KR. Vocal affect expression: A review and a model for future research. Psychological Bulletin 1986: 99: 2, 143-165.

Scherer KR. Vocal communication of emotion: A review of research paradigms. Speech Communication 2003: 40, 227-256.

Scherer KR, Wallbott HG. Evidence for universality and cultural variation of differential emotion response pattering. Journal of Personality and Social Psychology 1994: 66: 2, 310-328.

Schirmer A, Kotz SA. Sex differentiates the Stroop-effect in emotional speech: ERP evidence. In: Proceedings of the 1st Speech Prosody Conference. Aix-en-Provence, France. 2002, 631-634.

Schirmer A, Kotz SA, Friederici A. Sex differentiates the role of emotional prosody during word processing. Cogniotive Brain Research 2002: 14, 228-233.

Schirmer A, Simpson E. Brain correlates of vocal emotional processing in men and women. In: K. Izdebski (Ed.). Emotions in the Human Voice. Plural Publishing. San Diego. 2008, 75-86.

Schirmer A, Striano T, Friederici AD. Sex differences in the pre-attentive processing of vocal emotional expressions. Neuroreport 2005: 16: 6, 635-639.

Schröder M. Experimental study of affect bursts. Speech Communication 2003: 40: 1-2, 99-116.

Seppänen T, Väyrynen E. & Toivanen J. Prosody-based classification of emotions in spoken Finnish. Proceedings of the 8th European Conference on Speech Communication and Technology. Eurospeech, Geneva. 2003, 717-720.

Shipp T, Izdebski K. Vocal frequency and vertical larynx positioning by singers and nonsingers. Journal of The Acoustical Society of America. 1975: 58: 5.

Solomon R C. Philosophy of emotions. In E. Craig (Ed.) Routledge Encyclopedia of Philosophy. Vol. III. London & New York: Routledge. 1998, 285-290.

Van den Stock J, Righart R, de Gelder B. Body expressions influence recognition of emotions in the face and voice. Emotion 2007: 7: 3, 487-494.

Story BH, Laukkanen A-M, Titze IR. Acoustic impedance of an artificially lengthened and constricted vocal tract. Journal of Voice. 2000: 14: 4, 455-469.

Sundberg J. Singing and timbre. In: Music room acoustics. Royal Swedish Academy of Music No. 17. Stockholm. 1977, 57-81.

Sundberg J. The science of the singing voice. Northern Illinois University Press. DeKalb, Illinois. USA. 1987.

Sundberg J, Gauffin J. Waveform and spectrum of the glottal voice source. STL-QPSR. 1978, 35-50.

Švancara P, Horáček J, Vokřál J, Černý L. Computational modelling of effect of tonsillectomy on voice production. Logopedics, Phoniatrics, Vocology 2006: 31, 117-125.

Tartter VC, Baun D. Hearing smiles and frowns in normal and whisper registers. Journal of The Acoustical Society of America 1994: 96:4, 2101-2107.

Titze IR. Nonlinear source-filter coupling in phonation: Vocal exercises. Journal of The Acoustical Society of America 2008: 123: 4, 1902-1915.

Toivanen J, Seppänen T, Väyrynen E. Emotions in spoken Finnish: Cues for the human listener and computer. In: K. Izdebski (Ed.) Emotions in the Human Voice. Vol. I. Plural Publishing. San Diego. 2008, 101-108.

Tom K, Titze IR, Hoffman EA, Story BH. Three-dimensional vocal tract imaging and formant structure: Varying vocal register, pitch and loudness. Journal of The Acoustical Society of America 2001: 109:2, 742-747.

Vintturi J. Studies on voice production with a special interest on vocal loading, grender, some exposure factors and intensity regulation. Academic dissertation. The Medical Faculty of the University of Helsinki. Helsinki. 2001.

Waaramaa T, Laukkanen A-M. Gender differences in the perception of vocal expressions of emotions. European Communication Research and Education Association. 2nd European Communication Conference. Barcelona 25-28 November 2008. http://www.ecrea2008barcelona.org/guide/download/1164.pdf.

Waaramaa T, Laukkanen A-M, Alku P. Gender and expression of emotions. In: O'Dell ML, Nieminen T (toim./Eds.) Fontiikan päivät 2008. Tampere Studies in Language, Translation and Culture. Series B3. Tampere: Tampere University Press. Tampere. 2009, 65-72. In press. http://tampub.uta.fi/tup/978-951-44-7580-1.pdf.

Waaramaa T, Laukkanen A-M, Alku P, Björkner E, Leino T. Perception of emotions in mono-pitched vowels. In: Rantala L. (toim./ed.), Puheopin laitos. Raportteja 5/2007. Tampereen yliopisto. (Department of Speech Communication and Voice Research. Reports 5/2007. University of Tampere.)

Wambacq IJA, Shea-Miller KJ & Abubakr A. Non-voluntary and voluntary processing of emotional prosody: an event-related potentials study. Neuroreport 2004; 15: 3, 555-559.

Wayland R, Jongman A. Acoustic correlates of breathy and clear vowels: the case of Khmer. Journal of Phonetics 2003: 31: 2, 181-201.

Weaver JC, Anderson R J. Voice and personality interrelationships. The Southern Speech Communication Journal 1973: 38, 262-278.

Xu Y, Chuenwattanapranithi S. Perceiving anger and joy in speech through the size code. 16th International Congress of Phonetic Sciences (ICPHS). Saarbrücken, 6-10th August 2007.

Yu F, Chang E, Xu Y-Q, Shum H-Y. Emotion detection from speech to enrich multimedia content. Proceedings of the Second IEEE Pacific Rim Conference on Multimedia. Peking. 2001, 550-557.

Zei Pollermann B. A place for prosody in a unified model of cognition and emotion. In Speech prosody 2002, 17-22. Aix-en-Provence, France, April 11-13, 2002.

Zei Pollermann B, Izdebski K. A unified model of cognition, emotion, and action and its relation to vocally encoded cognitive-affective states. In: K. Izdebski (Ed.) Emotions in the Human Voice. Vol. I. Plural Publishing. San Diego. 2008, 43-64.

Perception of Emotional Valences and Activity Levels from Vowel Segments of Continuous Speech

*Teija Waaramaa, *Anne-Maria Laukkanen, †Matti Airas, and †Paavo Alku

Summary: This study aimed to investigate the role of voice source and formant frequencies in the perception of emotional valence and psychophysiological activity level from short vowel samples (~150 milliseconds). Nine professional actors (five males and four females) read a prose passage simulating joy, tenderness, sadness, anger, and a neutral emotional state. The stress carrying vowel [a:] was extracted from continuous speech during the Finnish word [t**a:**k:ahan] and analyzed for duration, fundamental frequency (F0), equivalent sound level (L_{eq}), alpha ratio, and formant frequencies F1–F4. Alpha ratio was calculated by subtracting the L_{eq} (dB) in the range 50 Hz–1 kHz from the L_{eq} in the range 1–5 kHz. The samples were inverse filtered by Iterative Adaptive Inverse Filtering and the estimates of the glottal flow obtained were parameterized with the normalized amplitude quotient ($NAQ = f_{AC}/(d_{peak}T)$). Fifty listeners (mean age 28.5 years) identified the emotional valences from the randomized samples. Multinomial Logistic Regression Analysis was used to study the interrelations of the parameters for perception. It appeared to be possible to identify valences from vowel samples of short duration (~150 milliseconds). NAQ tended to differentiate between the valences and activity levels perceived in both genders. Voice source may not only reflect variations of F0 and L_{eq}, but may also have an independent role in expression, reflecting phonation types. To some extent, formant frequencies appeared to be related to valence perception but no clear patterns could be identified. Coding of valence tends to be a complicated multiparameter phenomenon with wide individual variation.

Key Words: Voice quality - Perception of emotional valence - Inverse filtering.

INTRODUCTION

The communication of emotions plays a crucial role in human life. The two commonly studied dimensions, regarding both expression and perception of vocally conveyed emotions, are psychophysiological activity level (arousal) and valence.[1–3] Psychophysiological activity level is related to the emotion itself (eg anger vs. sadness), and it also reflects the strength or type of it (eg hot vs. cold anger). Valence refers to the affective value of the emotion: Neutrality, positivity, or negativity. It may be argued whether the human voice can be neutral in the sense of not including any kind of coloring in it or whether it can be perceived without any interpretation. Hence, neutrality has to be understood as a more or less hypothetical valence in the present study.

Studies on the vocal expression and perception of emotional content in speech have mainly concentrated on such prosodic characteristics as fundamental frequency (F0), sound pressure level (SPL), speech rate, and phoneme duration.[1,4,5] These characteristics have been found to be important discriminators in the production and perception of affective content in speech.[6–8] Fundamental frequency and variations in it have frequently been regarded as the main acoustic parameter of emotional information. In the communication of emotions, F0 may vary individually.[2] Typically it covaries with SPL. Both F0 and SPL increase in emotions with high activity level and decrease in low activity level emotions.[6,7]

The role of voice quality has received far less attention compared to F0, SPL, and duration as means of vocal communication. However, there is evidence that the more subtle emotive contents, such as affective strength and valence, are communicated by voice quality and also by rhythm.[1] In a narrow sense, the voice quality can be used to refer to any single vocal characteristic (such as pitch, register) which in turn may be situational, signaling, for instance, the speaker's emotional state, or a longer-term feature, for example, differentiating speakers or groups of speakers from each other. Laver defines voice quality "in a broad sense, as the characteristic auditory coloring of an individual speaker's voice."[9] Defined in this way, voice quality results from both phonatory and articulatory characteristics. It also interacts with other prosodic variables. Sometimes the terms "voice quality" and "timbre" have been used as synonyms. However, timbre mainly refers to the coloring of the voice on the axis darkness—brightness, while voice quality also comprises such glottal aspects as phonation type, turbulence noise and short-term temporal characteristics like perturbation and trembling. The term voice quality is used in this article due to its broader scope.

Phonatory characteristics can be studied from the point of view of variation in the glottal area[10] or the resulting glottal flow velocity waveform (the voice source).[11–15] To study voice source parameters in continuous speech, inverse filtering of the acoustic speech pressure signal recorded in free field is required.[15] The glottal waveform may be analyzed in the amplitude domain, for example, by calculating the normalized amplitude quotient (NAQ) which is the relative measure for the length of the glottal closing phase determined as a ratio of the flow amplitude to the negative peak of the first derivative of the flow waveform, divided by the period length. In an earlier study on NAQ, Airas and Alku[16] applied partly the same

Accepted for publication April 14, 2008.

Part of the paper was presented at The Voice Foundation's 35th Annual Symposium "Care of the Professional Voice," May 31–June 4, 2006, Philadelphia, and part of it at the "3rd World Voice Conference 2006," June 20–22, 2006, Istanbul.

From the *Department of Speech Communication and Voice Research, University of Tampere, Tampere, Finland; and the †Department of Signal Processing and Acoustics, Helsinki University of Technology, Espoo, Finland.

Corresponding author. Department of Speech Communication and Voice Research, University of Tampere, FIN-33014 Tampere, Finland. E-mail: teija.waaramaa@uta.fi

Journal of Voice, Vol. ■, No. ■, pp. 1-9

0892-1997/$34.00

doi:10.1016/j.jvoice.2008.04.004

material as in the present study. They reported females' tendency to use a wider scale and more extremes of emotional expressions than males.[16] Because the accuracy of inverse filtering technique is limited mainly to vowels with high F1, vowel [a:] was of interest in the present study.

NAQ has been found to differentiate between hypo- and hyperfunctional phonation (soft and strained voice quality) being high in the former and low in the latter.[15] Hypofunctional (soft or breathy) phonation is characterized by an almost sinusoidal glottal flow pulse shape whereas a relatively steep pulse shape is characteristic of hyperfunctional (strained or pressed) phonation type.[4,13] Hence, in hypofunctional phonation the open quotient (OQ = the open time of the glottis divided by the period length) is higher (and thus closed quotient, closed time divided by the period length, is lower) and speed quotient (the SQ opening time divided by the closing time of the glottis) is lower than in hyperfunctional phonation.[4,13] Flow pulse amplitude is typically lower in hyperfunctional phonation. Faster closing of the glottis or sufficiently high impedance of the vocal tract increases the glottal flow declination rate, thus, diminishing the spectral tilt. Hyperfunctional phonation has a gentler spectral slope than hypofunctional phonation.[13,17–20]

Glottal characteristics that may be related to phonation type are also known to covary with F0 and SPL.[4] In general, raising F0 increases open quotient, whereas higher intensity typically decreases it. Normally, SQ and closed quotient increase, and closed time divided (ClQ) decreases with higher intensity.[21–25] NAQ has been found to correlate well with SPL.[26] NAQ reflects phonation type by being high in hypofunctional voice and low in hyperfunctional voice.[15] According to Holmberg et al, the inverse filtered airflow waveform reflects more SPL than F0.[27] The articulatory characteristics affecting voice quality may also vary in relation to other prosodic variables. The length of the vocal tract may vary with F0[28,29]; raising the F0 typically leads to elevated laryngeal position and, thus, to a shortened vocal tract with higher formant frequencies. Formant frequencies, especially F1, are also used in intensity control: Increased mouth opening raises F1 closer to F2, which increases SPL.[30] Speech rate may affect voice quality parameters indirectly, for example, by increasing or decreasing the articulatory movements and, thus affecting formant frequencies.

Although voice quality interacts with other prosodic variables, it also tends to have a certain independence in communication. According to Ladd et al, F0 range, voice quality, and intonation contour type vary independently due to interspeaker and verbal context differences.[2] Gobl and Ní Chasaide also concluded that changes in voice quality alone may evoke widely differing associations in a verbally neutral expression.[31] Some studies have aimed to shed light on the individual role of the acoustic voice quality[7] in the expressions of emotions and their valence.[6,16,32–34] The results of Laukkanen et al suggested that voice source parameters may vary independently of F0 and intensity in emotional utterances.[4] In a further study by Laukkanen et al[6] the variation in F0, intensity and duration were artificially eliminated. The listeners appeared to categorize the samples of short duration (200 milliseconds) according to the psychophysiological activity level inherent in the emotions and the vocal effort level perceived in the samples. Perception of vocal effort seemed to be related to the glottal voice source waveform, and valence perception appeared to be related to F1 and F4. In a study by Waaramaa et al, the role of F0 was eliminated by having student actors express different emotional states on mono-pitched vowels [a:, o:, e:].[32] Intensity was allowed to vary freely in the production of the samples, but the SPL of the stimuli was normalized for the perceptual analysis. The results seemed to concur with those of Laukkanen et al: Psychophysiological activity level of the perceived emotions tended to be related to glottal voice source characteristics reflected in NAQ and alpha ratio (difference in level between the upper and lower frequencies in the spectrum).[6] High activity was related to a higher alpha ratio and a lower NAQ. Alpha ratio was calculated from the acoustic speech signal and thus it was also to some extent affected by formants. Formant frequencies tended to have relevance in valence perception, especially F3 and F4.[32] In another study by Waaramaa et al,[34] the role of F3 was investigated in the perception of valence of emotional expressions. Semisynthetic [a:] vowels with different F3 frequency values were used, the original F3, F3 lowered by 30 %, F3 raised by 30 % in frequency, and F3 totally removed. The results showed that the samples with a higher F3 value were perceived more often as positive than the original or the other manipulated samples. The results from the higher formant frequencies (F3 and F4) seem rational, because the two lower formants carry linguistic information, whereas the higher ones have more freedom to vary without disturbing the linguistic content and hence, may carry more information about nonlinguistic variables such as emotion related characteristics. The fact that smiling raises formant frequencies as it shortens the vocal tract, appears to explain the tendency that samples with higher F3 and F4 tended to be perceived as positive.

The present study investigated differences in voice quality parameters, that is, characteristics of the voice source and the formant frequencies, in emotional valence and activity perception. The focus was on the perception of short samples because short duration minimizes the effects of the prosodic variables, other than voice quality parameters, and thus, the effects of the voice source and filter may become more apparent. It is hypothesized that (1) vowel samples of short duration (~150 milliseconds) can be classified according to valence, even in cases where the identification of each actual emotion would not be possible,[35] (2) voice source characteristics, reflected in NAQ, are related to perception of valence and activity, (3) formant frequencies are related to perception of valence.

MATERIALS AND METHODS

Samples

The material was produced by nine professional actors, with healthy voices (five males and four females, age 26–45 years), who read aloud a prose extract both neutrally and expressing tenderness, sadness, anger, and joy. These emotional states were chosen because they represent both positive and negative valence and because they are likely to reflect different psychophysiological activity levels (anger and joy—higher activity—and

sadness and tenderness—lower activity). The duration of one text sample was approximately one minute. The main stress carrying vowel [a:] was extracted from the text from the Finnish word [t**a:**k:ahan] ('a burden indeed') for further analyses. The average duration of the vowel samples was 143 milliseconds.

The recordings were made in an anechoic room using a Sony DTC-690 DAT recorder and a Brüel & Kjær 4188 microphone at a distance of 50 cm from the subject's lips. Simulations of each emotional state were repeated ten times each in random order given by one of the experimenters. Thus, the collected data included 450 samples (9 actors × 5 emotions × 10 repetitions). This material was also used in earlier studies.[16,33,34] From the total number of 450 samples, 200 samples were randomly chosen for the analyses of the present study. It was considered that 200 samples would be the maximum in the perception test that the listeners could reasonably classify without becoming too tired. In 29 cases out of the 200 samples, the signal quality caused problems for automatic acoustic analyses (either due to too low signal-to-noise ratio or peak clipping) and these were excluded. The final number of samples was thus 171 (99 produced by males, 72 by females).

Perceptual analysis

The randomized [a:] vowel samples (N = 171) were replayed to 50 listeners (university teachers and students, 41 females, 9 males, mean age 28.5 years) for perception of the emotional valence and the psychophysiological activity level of the samples. The test included samples from every emotion category the actors had expressed. The identification of the emotions was not the aim of the listening test; the aim was rather to investigate which acoustic characteristics, if any, were in a statistically significant relation to emotional valence perceived. Four randomly formed groups evaluated the samples in a well-damped room. The listeners were seated approximately 2.5 m from the loudspeaker and the samples were replayed for them in random order at normal conversational loudness. A digital recorder and Genelec Biamp1019 A loudspeaker were used. The listeners completed a multiple-choice questionnaire which emotion they perceived. The results of the listening evaluation were studied by calculating the percentage of the most often chosen alternative (the maximum of the answers in each sample). The criterion was, however, that the maximum had to be above the chance level. Intrarater reliability was studied by calculating the percentage of the identical answers for those samples that were repeated (N = 19) in a random order in the course of the test.

Acoustic analyses

The [a:] vowels were extracted from the text samples using the SoundSwell Workstation (Hitech Development, Stockholm, Sweden). A total of 171 samples were analyzed for fundamental frequency (F0), equivalent sound level (L_{eq}), duration, alpha ratio, and formant frequencies F1–F4. Formant frequencies were measured with the aid of spectrograms and long-term average spectra taken from the middle of each vowel sample. Analyses were made with a signal analysis system Intelligent Speech Analyser developed by Raimo Toivonen, M.Sc. Eng. Alpha ratio

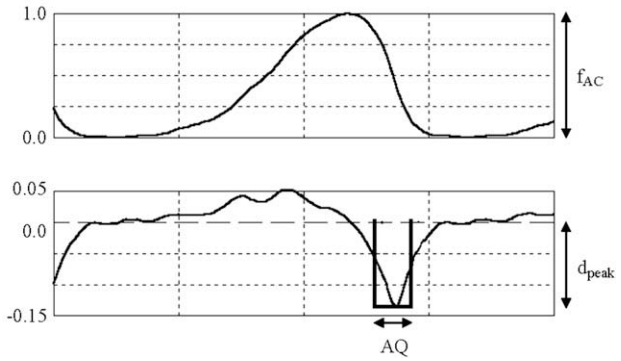


FIGURE 1. Schematic picture of an inverse filtered signal. Upper figure. Glottal flow: AC flow amplitude = f_{AC}. Lower figure. First derivative of glottal flow: Negative peak amplitude of the derivative = d_{peak}. AQ (amplitude quotient) = f_{AC}/d_{peak}. Flow on an arbitrary scale on the vertical axis, time on the horizontal axis.

estimates the spectral tilt originating from the glottal source by measuring the distribution of speech energy on the frequency axis and, thus, reflects the voice quality.[36] Here it was calculated by subtracting the L_{eq} (dB) in the range 50 Hz-1 kHz from the L_{eq} in the range 1–5 kHz.

The signal of the vowel samples was separated into the glottal flow (voice source) and the vocal tract filter functions[30] by inverse filtering it with the Iterative Adaptive Inverse Filtering method.[37] Thereafter, the voice source was parameterized by calculating the NAQ ($NAQ = f_{AC}/(d_{peak}T)$)[15] for 40-millisecond portion of the samples. To determine NAQ, the peak-to-peak AC flow (f_{AC}) and the absolute amplitude of the negative peak of the first derivative of the flow waveform (d_{peak}) were measured. T is the fundamental period length. NAQ is used to measure the relative time of the glottal closing phase. It has been shown to be robust against artefacts such as formant ripple present in the glottal flow estimates (Figure 1).[15,16,38]

Statistical analyses

Mean values of F0, L_{eq}, duration, NAQ, alpha ratio, and formant frequencies F1–F4 of the samples were calculated. Relations between the acoustic variables were investigated with bivariate Pearson correlation coefficients. Multinomial Logistic Regression Analysis was used to study the relations of acoustic variables to the perceived valence and psychophysiological activity level. Valence was marked by values: −1 = negative, 0 = neutral, +1 = positive valence, and activity was marked as −1 = low, 0 = intermediate, +1 = high psychophysiological activity level. Statistical analyses were carried out using SPSS-15 software (SPSS Inc., Chicago, IL).

RESULTS

The degree of agreement among the listeners in the perception of the actual emotional states was 60%. Intrarater reliability in percentages was 58 for the *emotional states* perceived and 61 for the *valence* perceived. The aim was not to study the correct identification of the samples, but rather the acoustic characteristics of the samples that had been perceived as reflecting

TABLE 1.
Means, standard deviations, minima and maxima of acoustic parameters F0, L_{eq}, duration, alpha ratio, NAQ, and formant frequencies F1–F4 in the emotions perceived by the listeners

	Emotion perceived	Males					Females				
		N	Mean	S.D.	Min	Max	N	Mean	S.D.	Min	Max
L_{eq} (dB)	Neutral	40	68	5	60	80	17	67	3	63	74
	Sadness	33	65	6	57	78	16	62	5	52	74
	Joy	3	70	5	64	74	14	74	5	63	82
	Anger	13	79	2	76	84	19	74	5	63	84
	Tenderness	10	63	3	59	67	6	60	2	58	63
	Total	99	68	7	57	84	72	69	7	52	84
Duration (millisecond)	Neutral	40	150	30	80	227	17	124	26	87	170
	Sadness	33	151	29	102	210	16	122	29	76	187
	Joy	3	162	32	125	182	14	141	43	85	246
	Anger	13	164	25	123	208	19	136	32	83	187
	Tenderness	10	158	38	104	246	6	163	45	107	227
	Total	99	153	30	80	246	72	133	35	76	246
Alpha ratio (dB)	Neutral	40	−6	2	−10	−2	17	−6	4	−12	0
	Sadness	33	−8	3	−16	0	16	−8	4	−18	2
	Joy	3	−7	3	−10	−4	14	0	5	−8	11
	Anger	13	−4	4	−10	3	19	1	3	−7	6
	Tenderness	10	−8	5	−16	−1	6	−11	4	−16	−5
	Total	99	−6	3	−16	3	72	−4	6	−18	11
NAQ	Neutral	37	0.09	0.02	0.06	0.15	15	0.10	0.02	0.06	0.15
	Sadness	25	0.11	0.03	0.07	0.17	14	0.14	0.03	0.07	0.18
	Joy	2	0.12	0.01	0.12	0.13	12	0.12	0.04	0.06	0.23
	Anger	12	0.08	0.02	0.06	0.12	14	0.10	0.04	0.06	0.20
	Tenderness	6	0.14	0.04	0.09	0.18	5	0.16	0.03	0.13	0.21
	Total	82	0.10	0.03	0.06	0.18	60	0.12	0.04	0.06	0.23
F0 (Hz)	Neutral	40	141	41	86	258	17	215	34	172	258
	Sadness	33	138	48	86	258	16	272	68	172	474
	Joy	3	186	25	172	215	14	320	98	172	517
	Anger	13	202	32	172	258	19	288	93	172	474
	Tenderness	10	120	34	86	172	6	201	59	129	301
	Total	99	147	47	86	258	72	266	85	129	517
F1* (Hz)	Neutral	40	647	33	603	732	17	681	44	603	775
	Sadness	33	629	56	517	732	16	662	98	603	947
	Joy	3	646	43	603	689	14	747	112	603	1034
	Anger	13	686	69	603	775	18	706	87	603	904
	Tenderness	10	655	53	560	732	6	675	104	603	861
	Total	99	647	51	517	775	71	696	91	603	1034
F2* (Hz)	Neutral	40	1242	62	1163	1421	17	1363	87	1249	1550
	Sadness	33	1280	69	1163	1421	16	1359	182	1076	1809
	Joy	3	1249	86	1163	1335	14	1375	190	1034	1593
	Anger	13	1249	56	1120	1335	18	1340	163	861	1636
	Tenderness	10	1301	86	1206	1464	6	1414	157	1292	1637
	Total	99	1262	69	1120	1464	71	1363	156	861	1809
F3* (Hz)	Neutral	40	2612	190	2326	3101	17	2749	255	2326	3187
	Sadness	33	2637	139	2411	2972	16	2756	505	1292	3402
	Joy	3	2713	114	2584	2799	14	2938	593	1809	3919
	Anger	13	2726	118	2498	2885	18	2792	490	1507	3574
	Tenderness	10	2743	157	2541	3015	6	3000	192	2713	3230
	Total	99	2652	165	2326	3101	71	2820	451	1292	3919
F4† (Hz)	Neutral	39	3732	330	3273	4565	17	4200	338	3144	4608
	Sadness	33	3761	396	3015	5039	16	4153	372	3359	4608
	Joy	3	3603	174	3445	3790	14	4119	498	3230	4780
	Anger	13	3687	257	3402	4264	18	4118	408	3230	4737
	Tenderness	10	3781	423	3359	4737	6	4436	234	4134	4780
	Total	98	3737	348	3015	5039	71	4172	392	3144	4780

* One outlier excluded in females.
† One outlier excluded in both genders.

certain emotions with an agreement above the chance level. This was the case for all samples. The samples were grouped according to the answers of the majority of listeners. The total numbers of samples perceived as representing certain emotions were neutral = 57, sadness = 49, anger = 32, joy = 17, and tenderness = 16. Out of the 99 samples from male subjects, 40 were perceived as neutral, 33 as expressions of sadness, 13 as anger, 10 as tenderness, and 3 as joy. Out of the 72 samples from females, the distribution was 17 for neutral, 19 for anger, 16 for sadness, 14 for joy, and 6 for tenderness.

When valence was derived from the answers of the listeners, the numbers were as follows: Out of the samples produced by males, 46 (46.5%), were perceived as negative, 40 (40.4%) perceived as neutral and 13 (13.1%) perceived as positive; samples produced by females, 35 (48.6%) perceived as negative, 17 (23.6%) perceived as neutral and 20 (27.8%) perceived as positive.

Descriptive statistics of F0, L_{eq}, duration, NAQ, alpha ratio, and formant frequencies F1–F4 of the samples can be seen in Table 1. The value of NAQ was missing in the case of 17 male and 12 female samples due to the too short duration and sections without stable periods. Bivariate Pearson's correlation coefficients of F0, L_{eq}, duration, NAQ, alpha ratio, and formant frequencies F1–F4 of the samples are shown in Table 2.

Table 2 shows the results of Pearson correlations separately for males and females. In both genders, NAQ correlated with F0 and $L_{eq,}$ and alpha ratio correlated with NAQ. Furthermore, F0, L_{eq}, alpha ratio, and F1 correlated pair wise in both genders.

Valence. In the Multinomial Logistic Regression Analysis, valence was set as dependent variable. Neutrality was used as the reference category. Perception of valence was studied separately for both genders. The results can be seen in Table 3. Because of the large number of missing values in NAQ the analyses were made with NAQ (on the right-hand side) and without NAQ (on the left-hand side). Here L_{eq} did not get any significant results in neither of the genders because soft and loud samples were included in both positive and negative valences.

In *males*, F3 frequency had a significant effect on valence perceived ($P = 0.002$) when NAQ was excluded from the analyses, higher values of F3 indicating a higher probability of positive than negative valence. When NAQ was included, duration appeared to be associated with valence being longer in positive valence. However, in males anger (N = 13) had the longest duration (on the average 164 milliseconds) of the emotions studied. Hence, the statistical result for the relation between duration and valence may be emphasized by the high number of the sadness samples (N = 33) and their shorter duration (on the average 151 milliseconds). NAQ seemed to have a significant effect on the perception of valence ($P = 0.048$). Small value of NAQ and short duration indicated higher probability of perceiving neutral valence in males (Figure 2A).

In *females*, F0 had significance in the valence perception with and without NAQ ($P = 0.032$ and $P = 0.004$), and F1 when NAQ was included ($P = 0.043$). NAQ had a significant effect in the perception of valence, smaller value of NAQ indicating a higher probability of neutral valence ($P = 0.012$) (Figure 2B).

To reduce the effects of individual differences in formant frequencies and to identify possible tendencies, the mean differences in the formant frequencies were calculated in percentages between the emotional states and compared to the neutral valence, which was given a hypothetic value of 0. The results can be seen in Figure 3A and B.

The tendency for F3 to reach a higher frequency in positive valences than in negative ones is seen in the Figure 3A and B. However, the unequal number of some sample groups, especially in males, may obscure the tendencies. In females, the tendency is clearer.

TABLE 2.
Pearson correlations between the acoustic variables studied in males and females, separately

		Duration	Alpha ratio	NAQ	F0	F1	F2	F3	F4*
Males	L_{eq}	0.250†	0.475‡	−0.579‡	0.701‡	0.354‡	−0.162	0.257†	0.102
	Duration		0.133	−0.146	0.278‡	0.079	−0.121	0.096	0.152
	Alpha ratio			−0.510‡	0.490‡	0.483‡	0.073	−0.042	0.012
	NAQ				−0.484‡	−0.109	0.218†	0.191	0.185
	F0					0.227†	0.231†	0.063	−0.099
	F1						−0.243†	0.229†	0.024
	F2							−0.112	−0.093
	F3								0.520‡
Females	L_{eq}	0.071	0.776‡	−0.320†	0.424‡	0.423‡	−0.031	0.161	0.034
	Duration		0.057	0.048	−0.032	0.083	0.143	0.033	−0.022
	Alpha ratio			−0.309†	0.553‡	0.523‡	0.111	0.130	−0.057
	NAQ				0.268†	0.112	0.091	0.211	0.156
	F0					0.564‡	0.183	0.446‡	0.243†
	F1†,§						0.194	0.092	−0.033
	F2†,§							0.409‡	0.227
	F3†,§								0.793‡

* One outlier excluded in both genders.
† Correlation is significant at the 0.05 level (two tailed).
‡ Correlation is significant at the 0.01 level (two tailed).
§ One outlier excluded in females.

TABLE 3.
Results of Multinomial Logistic Regression Analysis on the valence perceived: effects of L_{eq}, duration, alpha ratio, NAQ, and formant frequencies

	Full model without NAQ			Full model		
	P-value*	Negative valence; OR (95 % CI)	Positive valence; OR (95 % CI)	*P*-value*	Negative valence; OR (95 % CI)	Positive valence; OR (95 % CI)
Males	(n = 98)			(n = 81)		
Leq	0.134	1.04 (0.91–1.18)	0.86 (0.70–1.06)	0.312	1.12 (0.95–1.31)	0.96 (0.69–1.35)
Duration†	0.089	1.08 (0.91–1.29)	1.34 (1.02–1.75)	0.004	1.15 (0.94–1.42)	1.95 (1.16–3.27)
Alpharatio	0.585	0.90 (0.75–1.10)	0.95 (0.72–1.26)	0.819	0.93 (0.73–1.18)	0.97 (0.58–1.64)
F0†	0.835	1.01 (0.84–1.20)	0.93 (0.70–1.23)	0.990	1.01 (0.81–1.26)	1.03 (0.67–1.58)
F1†	0.659	1.01 (0.90–1.14)	1.09 (0.90–1.31)	0.710	0.95 (0.83–1.09)	0.99 (0.79–1.25)
F2†	0.013	1.11 (1.01–1.21)	1.20 (1.03–1.38)	0.068	1.12 (1.01–1.24)	1.09 (0.91–1.31)
F3†	0.002	1.03 (0.99–1.07)	1.13 (1.04–1.23)	0.088	1.03 (0.98–1.08)	1.16 (0.98–1.38)
F4†	0.132	0.99 (0.98–1.01)	0.97 (0.94–1.00)	0.048	0.99 (0.97–1.01)	0.92 (0.84–1.01)
NAQ‡				0.003	1.33 (0.97–1.83)	2.49 (1.27–4.87)
Females	(n = 71)			(n = 60)		
L_{eq}	0.692	0.96 (0.80–1.14)	1.02 (0.83–1.25)	0.246	1.04 (0.79–1.36)	1.19 (0.89–1.58)
Duration†	0.070	1.10 (0.88–1.37)	1.26 (1.01–1.58)	0.247	1.05 (0.78–1.40)	1.20 (0.91–1.59)
Alpharatio	0.435	1.11 (0.87–1.40)	0.98 (0.76–1.28)	0.164	1.27 (0.92–1.77)	1.05 (0.74–1.49)
F0†	0.004	1.25 (1.07–1.45)	1.15 (0.98–1.35)	0.032	1.26 (1.02–1.55)	1.12 (0.90–1.40)
F1†	0.089	0.91 (0.79–1.05)	0.99 (0.86–1.14)	0.043	0.85 (0.70–1.00)	0.95 (0.78–1.15)
F2†	0.897	0.99 (0.94–1.04)	0.99 (0.93–1.05)	0.605	0.97 (0.91–1.04)	0.96 (0.90–1.04)
F3†	0.148	1.00 (0.97–1.03)	1.03 (0.99–1.06)	0.116	1.01 (0.97–1.05)	1.04 (0.99–1.08)
F4†	0.350	0.99 (0.96–1.02)	0.98 (0.95–1.01)	0.121	0.99 (0.96–1.03)	0.96 (0.93–1.00)
NAQ‡				0.012	1.56 (1.08–2.26)	1.61 (1.08–2.39)

Note: The reference category is "neutral."

* Significance of the effect in the likelihood ratio test between final model and a reduced model. OR (Odds Ratio) implies the probability of the event compared to the reference group per difference of one unit in the explanatory variable. OR of 1 implies that the event is equally likely. OR greater than 1 implies that the event is more likely in the examined group. OR less than 1 implies that the event is less likely in the examined group.

† Values of duration and F0–F4 are divided by 10 to make the ORs interpretable. Thus, OR implies the probability of the event compared to the reference group per difference of 10 unit in duration or F0–F4.

‡ Values of NAQ are multiplied by 100. Thus, OR implies the probability of the event compared to the reference group per difference of 0.01 unit in NAQ.

Psychophysiological activity level. The results of the Multinomial Logistic Regression Analysis for psychophysiological activity level are seen in Table 4. With and without NAQ, perception of activity level was mainly associated with L_{eq} in both genders ($P < 0.001$). When NAQ was excluded in *males*, formant frequencies F2–F4 appeared have significance in perception of the activity level (F2: $P = 0.017$, F3: $P = 0.027$, F4: $P = 0.0017$). The frequency of F4 seemed to be lowest in the samples with high activity level (joy and anger). When NAQ was included F2 and F4 remained significant (F2: $P = 0.006$, F4: $P = 0.008$) and additionally, alpha ratio tended to have significance in the perception of the activity level ($P = 0.042$). The frequency of F2 was lowest for the medium activity level. NAQ had a significant effect in the perception of the activity levels ($P = 0.013$). The value of NAQ was greatest for low activity level. In *females* besides L_{eq}, F0 appeared to have significance both with ($P = 0.017$) and without ($P = 0.003$) NAQ. When NAQ was included perception of the activity level appeared to be coded through F4 ($P = 0.022$). The difference in NAQ results between the genders may occur due to the unequal number of the samples in the different groups of the activity levels.

DISCUSSION

Short vowel [a:] samples (average 143 milliseconds) were used as material to identify the very essential characteristics used in the perception and categorization of emotions. By using short samples, it was possible to eliminate the effects of such prosodic characteristics as an overall pitch envelope, which has been found to be a main source in conveying emotional information in longer utterances.[1] The elimination of those prosodic characteristics, in turn, was considered important in identifying the role of voice quality alone in the emotional communication. On the other hand, a question may arise whether the samples used in the present study were long enough in duration in order for any emotional information to be perceived. According to earlier findings, the recognition of emotional information takes place within the first 100–150 milliseconds of the expression and tends to be based primarily on voice quality.[35] Based on neural processes, the perception of valence is even faster than the cognitive identification of an actual emotion.[39,40] According to Scherer,[41] emotional vocal samples have been recognized in earlier studies with approximately 60% accuracy. The recognition of emotions from facial expressions is somewhat higher.[41]

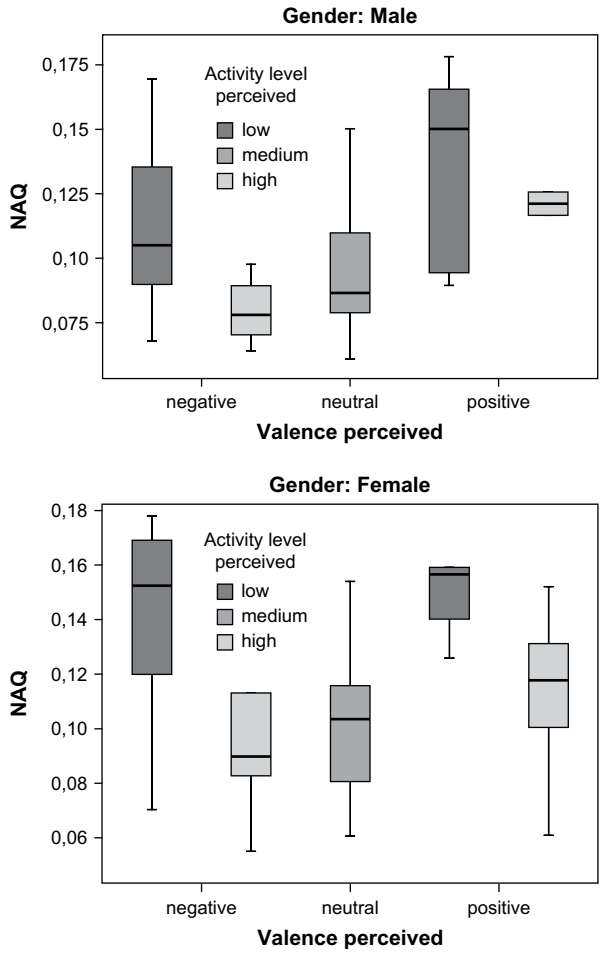


FIGURE 2. **A** and **B:** Boxplot graphs of the NAQ values in the valences perceived, clusters defined by the activity levels.

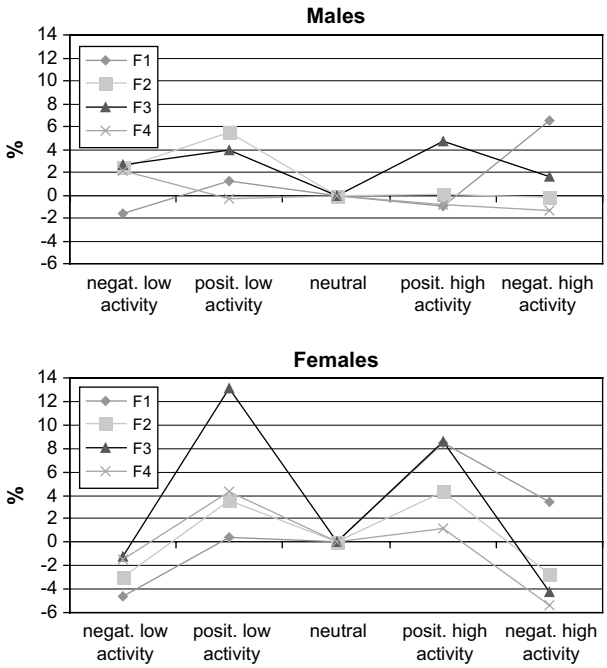


FIGURE 3. **A** and **B:** Graphs of the percentual differences in formant frequencies in valences perceived. Neutral valence was set as a reference category and given a hypothetic value of 0.

Naturally, samples with longer duration include more indicators of the emotional state of the speaker than one single vowel. In the present study, the focus was on the perception of emotional characteristics and not on the identification of actual emotions. The listeners agreed with each other and were consistent in repeated evaluation clearly above chance level. Therefore, the samples used in the current study appeared to be adequate in quality and duration.

Emotional samples produced by actors have been claimed to be acted, not "real" emotional expressions. They have also been claimed to be stereotypical and controlled. However, it may be argued that if the emotional expressions by actors were not recognizable, it would make acting impossible.[41] The audience needs to be given some hints about the emotional states expressed. If the expressions were completely individual, and thus, generally irregular, they would not be recognized by the audience who does not know the performer personally. It can also be questioned if our "real" expressions of emotions are not controlled because there are several ways in which our social behavior is controlled, for example, by the environment, authorities, or other social systems.[7,42] Therefore one might ask if there are any "real" emotional expressions in our social lives or whether they occur only in totally uncontrolled circumstances. Also in "real" life, personality and social competence are factors that determine our skills in emotional behavior. As a result, it seemed reasonable to use acted expressions of emotions as the material of the present study.

The results of the present study appear to suggest that valence was differentiated from neutrality mainly by NAQ in both genders, being somewhat lower in neutral valence than in other valences. Also some earlier studies support this result suggesting that NAQ may have an effect on the perception of valence.[32] A certain caution, however, is needed in the interpretation of the results of NAQ in relation to valence, because activity may confuse the results because of the inequality of the number of the different emotion samples. NAQ was greater, for example, in tenderness than in joy and analogously greater in sadness than in anger, and yet both tenderness and joy represent positive valence and sadness and anger negative valence (see Figure 2A and B).

Perception of valence from *females*' samples was associated besides NAQ also with F0 and F1. On the average the frequency of F0 was lowest in neutrality. In the samples with high effort level the frequency of F0 was higher in positive than in negative valence. The situation was the opposite in the samples with low effort level, F0 being lower in positive valence than in negative valence. In females, frequency of F1 was higher in positive than in negative valence, which may imply a larger mouth opening and hence strengthening the audibility of F1 by moving it closer to F2.[30] (Also frequency of F2 was higher in positive valence than in negative one.) Higher F1 frequency may also imply shortening of the vocal tract by a smiling position. In *both genders*, a correlation was found between L_{eq} and F0 and F1. F0 and F1 correlation may imply that the height of the larynx varies with F0, and F0 is known to covary with L_{eq}.

TABLE 4.
Results of Multinomial Logistic Regression Analysis on the psycho-physiological activity level perceived: effects of L_{eq}, duration, alpha ratio, NAQ, and formant frequencies

	Full model without NAQ			Full model		
	P-value*	Low activity; OR (95 % CI)	High activity; OR (95 % CI)	*P*-value*	Low activity; OR (95 % CI)	High activity; OR (95 % CI)
Males	(n = 98)			(n = 81)		
Leq	0.001	0.90 (0.78–1.03)	1.46 (1.09–1.97)	<0.001	0.87 (0.71–1.06)	1.90 (1.20–3.03)
Duration†	0.287	1.14 (0.95–1.38)	1.16 (0.85–1.59)	0.088	1.30 (1.01–1.68)	1.09 (0.68–1.74)
Alpharatio	0.383	0.88 (0.71–1.09)	1.15 (0.71–1.86)	0.042	0.73 (0.50–1.06)	1.67 (0.74–3.78)
F0†	0.965	0.98 (0.81–1.18)	0.97 (0.70–1.36)	0.633	0.95 (0.74–1.22)	1.23 (0.70–2.16)
F1†	0.927	1.02 (0.89–1.15)	0.96 (0.71–1.29)	0.456	0.94 (0.80–1.11)	0.83 (0.59–1.16)
F2†	0.017	1.14 (1.03–1.25)	1.05 (0.86–1.29)	0.006	1.21 (1.06–1.37)	1.08 (0.83–1.41)
F3†	0.027	1.04 (1.00–1.08)	1.10 (1.00–1.22)	0.151	1.03 (0.98–1.09)	1.10 (0.97–1.26)
F4†	0.017	1.00 (0.98–1.01)	0.94 (0.89–0.99)	0.008	1.00 (0.97–1.02)	0.92 (0.85–0.99)
NAQ‡				0.013	1.03 (0.72–1.47)	2.93 (1.11–7.71)
Females	(n = 71)			(n = 60)		
L_{eq}	<0.001	0.66 (0.46–0.94)	1.33 (1.00–1.77)	<0.001	0.22 (0.04–1.26)	2.53 (1.02–6.25)
Duration†	0.125	1.23 (0.88–1.71)	1.31 (0.94–1.82)	0.797	0.83 (0.37–1.84)	1.15 (0.57–2.33)
Alpharatio	0.014	0.84 (0.61–1.17)	1.49 (1.01–2.20)	0.493	1.25 (0.52–3.02)	1.52 (0.71–3.22)
F0†	0.003	1.39 (1.07–1.80)	1.21 (0.99–1.48)	0.017	1.53 (0.87–2.69)	1.51 (0.93–2.44)
F1†	0.605	1.00 (0.82–1.23)	0.91 (0.74–1.12)	0.104	1.27 (0.87–1.85)	0.74 (0.47–1.19)
F2†	0.955	0.99 (0.91–1.08)	0.99 (0.90–1.08)	0.419	0.86 (0.64–1.15)	1.04 (0.82–1.32)
F3†	0.318	1.00 (0.96–1.05)	1.04 (0.98–1.10)	0.404	1.07 (0.93–1.24)	1.02 (0.90–1.15)
F4†	0.106	0.99 (0.94–1.03)	0.95 (0.90–1.00)	0.022	1.00 (0.87–1.16)	0.90 (0.81–1.01)
NAQ‡				0.093	1.97 (0.82–4.69)	1.75 (0.72–4.22)

Note: The reference category is "medium."

* Significance of the effect in the likelihood ratio test between final model and a reduced model. OR (Odds Ratio) implies the probability of the event compared to the reference group per difference of one unit in the explanatory variable. OR of 1 implies that the event is equally likely. OR greater than 1 implies that the event is more likely in the examined group. OR less than 1 implies that the event is less likely in the examined group.

† Values of duration and F0–F4 are divided by 10 to make the ORs interpretable. Thus, OR implies the probability of the event compared to the reference group per difference of 10 unit in duration or F0–F4.

‡ Values of NAQ are multiplied by 100. Thus, OR implies the probability of the event compared to the reference group per difference of 0.01 unit in NAQ.

Consequently, a correlation between L_{eq} and F0 and filter functions was to be expected. Higher F1 as such also tends to raise L_{eq}, because it is brought closer to F2. A decrease in the frequency distance between formants increases their amplitudes and enhances their loudness. However, it might be argued that higher F0 frequency in females lowers the reliability of formant detection, thereby impairing the inverse filtering of the females' samples. This may naturally concern male samples with high F0 as well.

The role of formant frequencies in the valence perception appeared to be smaller than expected in this material. Hence, the results of the formant frequencies did not support the hypothesis concerning their role in emotion perception. Their role may differ between activity levels and furthermore, individual differences vary over a wide scale. Thus, valence coding tends to depend on several parameters. Other parameters studied appeared to carry more emotional information than formant frequencies, apparently due to the strong linguistic task of the formants. This warrants further study.

It has been difficult to show statistically any certain pattern of formant frequencies through which emotional valence was coded. However, there appeared to be a tendency for positive valence perceived to have somewhat higher F3 frequency than in neutral or negative valences. This may be partly due to individual differences and partly to a wide range of means trying to code valence in different activity levels. The linguistic importance of formants may disturb any systematic use of them in emotional communication. Other characteristics, for example, the spectral tilt (reflected in intensity) and distance between the partials (connected to F0) may also have an effect on formant realization.

CONCLUSIONS

1 It appeared to be possible to perceive emotional valence from vowel samples of ~150 milliseconds in duration.
2 NAQ was lower in neutrality than in the other valences in both genders, which mainly suggests interaction between psychophysiological activity level and valence in the present material, for example, due to differences in the number of samples representing different emotions.
3 Psychophysiological activity level was coded mainly through L_{eq} in both genders.

4 Perception of positive valence tended to be related to higher F3 but no clear pattern could be identified, for example, due to differences in formant use on different activity levels.
5 Perception of valence tends to be a complex multiparameter task with wide interindividual variation.

Acknowledgments

This study was supported by the Academy of Finland (grants no 200807 and no 200859). Hanna-Mari Pasanen, M.Sc., from the Unit for Science, Technology and Innovation Studies (TaSTI) University of Tampere, is thanked for statistical analyses.

REFERENCES

1. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J Acoust Soc Am*. 1993;93(2):1097–1108.
2. Ladd DR, Silverman KEA, Tolkmitt F, Bergmann G, Scherer KR. Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *J Acoust Soc Am*. 1985;78(2):435–444.
3. Cowie R, Cornelius RR. Describing the emotional states that are expressed in speech. *Speech Commun*. 2003;40(1–2):1–33.
4. Laukkanen A-M, Vilkman E, Alku P, Oksanen H. Physical variations related to stress and emotional state: a preliminary study. *J Phonet*. 1996;24:313–335.
5. Lieberman P, Michaels SB. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *J Acoust Soc Am*. 1962;34(7):922–927.
6. Laukkanen A-M, Vilkman E, Alku P, Oksanen H. On the perception of emotions in speech: the role of voice quality. *Log Phon Voc*. 1997;22(4):157–168.
7. Banse R, Scherer KR. Acoustic Profiles in Vocal Emotion Expression. *J Personality and Soc Psychol*. 1996;70(3):614–636.
8. Scherer KR. Vocal communication of emotion: a review of research paradigms. *Speech Commun*. 2003;40:227–256.
9. Laver J. *The phonetic description of voice quality*. Great Britain: Cambridge University Press; 1980.
10. Granqvist S, Hertegård S, Larsson H, Sundberg J. Simultaneous analysis of vocal vibration and transglottal airflow; exploring a new experimental set-up. *STL-QPSR*. 2003;45:35–46.
11. Rothenberg M. Some relations between glottal air flow and vocal fold contact area. Accessed July 5, 2006. Available at http://www.rothenberg.org/vfca/vfca.htm.
12. Sundberg J, Gauffin J. Waveform and spectrum of the glottal voice source. *STL-QPSR* 1978;35–50.
13. Gauffin J, Sundberg J. Spectral correlates of glottal voice source waveform characteristics. *J Speech and Hear Res*. 1989;32:556–565.
14. Cummings KE, Clementes MA. Analysis of the glottal excitation of emotionally styled and stressed speech. *J Acoust Soc Am*. 1995;98(1):88–98.
15. Alku P, Bäckström T, Vilkman E. Normalized amplitude quotient for parametrization of the glottal flow. *J Acoust Soc Am*. 2002;112(2):701–710.
16. Airas M, Alku P. Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*. 2006;63:26–46.
17. Fant G, Lin Q. Glottal source-vocal tract acoustic interaction. *STL-QPSR*. 1987;1:13–27.
18. Granström B, Nord L. Ways of exploring speaker characteristics and speaking styles. *Proceedings of the 12th International Congress of Phonetic Sciences* 1991;278–281. 19–24 August Aix-en-Provence.
19. Laukkanen A-M. On speaking voice exercises. A study on the acoustic and physiological effects of speaking voice exercises applying manipulation of the acoustic-aerodynamic state of the supraglottic space and artificially modified auditory feedback. Doctoral dissertation, . *Medical School*. Finland: University of Tampere; 1995.
20. Alku P, Vilkman E, Laukkanen A-M. Estimation of amplitude features of the glottal flow by inverse filtering speech pressure signals. *Speech Commun*. 1998;24:123–132.
21. Sonesson B. On the anatomy and vibratory pattern of the human vocal folds. With special reference to a photo-electrical method for studying the vibratory movements. In: *Acta Oto-laryngologica*, supplementum 156, . *Sweden: Department of Anatomy and Department of Otolaryngology*. Lund: University of Lund; 1960.
22. Sundberg J, Andersson M, Hultqvist C. Effects of subglottal pressure variation on professional baritone singers' voice sources. *J Acoust Soc Am*. 1999;105:1965–1971.
23. Sundberg J, Fahlstedt E, Morell A. Effects on the glottal voice source of vocal loudness variation in untrained female and male voices. *J Acoust Soc Am*. 2005;117:879–885.
24. Sundberg J, Titze I, Scherer R. Phonatory control in male singing: a study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source. *J Voice*. 1993;7:15–29.
25. Holmberg EB, Hillman RE, Perkell JS. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *J Acoust Soc Am*. 1988;84(2):511–529.
26. Vilkman E, Alku P, Vintturi J. Dynamic extremes of voice in the light of time domain parameters extracted from the amplitude features of glottal flow and its derivative. *Folia Phoniatr logop*. 2002;54:144–157.
27. Holmberg EB, Hillman RE, Perkell JS. Glottal airflow and transglottal air pressure measurements for male and female speakers in low, normal, and high pitch. *J Voice*. 1989;3(4):294–305.
28. Shipp T, Izdebski K. Vocal frequency and vertical larynx positioning by singers and nonsingers. *J Acoust Soc Am*. 1975;58:5.
29. Story BH, Laukkanen A-M, Titze IR. Acoustic impedance of an artificially lengthened and constricted vocal tract. *J Voice*. 2000;14(4):455–469.
30. Fant G. *Acoustic Theory of Speech Production. With Calculations Based on X-Ray Studies of Russian Articulations*. 2nd ed. The Hague: Mouton; 1970.
31. Gobl C, Ní Chasaide A. The role of voice quality in communicating emotion, mood and attitude. *Speech Commun*. 2003;40(1–2):189–212.
32. Waaramaa T, Laukkanen A-M, Alku P, Björkner E, Leino T. Perception of emotions in mono-pitched vowels. In: Rantala L. (toim./ed.), Puheopin laitos. Raportteja 5/2007. Tampereen yliopisto. (Department of Speech Communication and Voice Research. Reports 5/2007. University of Tampere.)
33. Toivanen J, Waaramaa T, Alku P, Laukkanen A-M, Seppänen T, Väyrynen E, Airas M. Emotions in /a:/: a perceptual and acoustic study. *Log Phon Voc*. 2006;31(1):43–48.
34. Waaramaa T, Alku P, Laukkanen A-M. The role of F3 in the vocal expression of emotions. *Log Phon Voc*. 2006;31(4):153–156.
35. Bostanov V, Kotchoubey B. Recognition of affective prosody: continuous wavelet measures of event-related brain potentials to emotional exclamations. *Psychophysiology*. 2004;41:259–268.
36. Frøkjær-Jensen B, Prytz S. Registration of voice quality. *Brüel & Kjær Technical Review*. 1973;3:3–17.
37. Alku P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun*. 1992;11(2–3):109–118.
38. Bäckström T, Alku P, Vilkman E. Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range. *IEEE Trans Speech Audio Process*. 2002;10(3):186–192.
39. Damasio A. Looking for Spinoza: Joy, Sorrow, and the Feeling Brain. *A Harvest Book, Harcourt, Inc. USA* 2003.
40. Zei Pollermann B. A place for prosody in a unified model of cognition and emotion. *Speech Prosody 2002. Aix-en-Provence, France* April 11–13, 2002; SP-2002: 17–22.
41. Scherer KR. Vocal communication of emotion: a review of research paradigms. *Speech Commun*. 2003;40:227–256.
42. Feldman Barrett L. Are emotions natural kind? *Perspectives on Psychol Sci*. 2006;1:1.

Logopedics Phoniatrics Vocology. 2006; 31: 43–48

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Emotions in [a]: A perceptual and acoustic study

JUHANI TOIVANEN[1], TEIJA WAARAMAA[2], PAAVO ALKU[3], ANNE-MARIA LAUKKANEN[2], TAPIO SEPPÄNEN[4], EERO VÄYRYNEN[4] & MATTI AIRAS[3]

[1]*MediaTeam, University of Oulu and Academy of Finland,* [2]*Department of Speech Communication and Voice Research, University of Tampere, Finland,* [3]*Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland,* [4]*MediaTeam, University of Oulu, Finland*

Abstract
The aim of this investigation is to study how well voice quality conveys emotional content that can be discriminated by human listeners and the computer. The speech data were produced by nine professional actors (four women, five men). The speakers simulated the following basic emotions in a unit consisting of a vowel extracted from running Finnish speech: neutral, sadness, joy, anger, and tenderness. The automatic discrimination was clearly more successful than human emotion recognition. Human listeners thus apparently need longer speech samples than vowel-length units for reliable emotion discrimination than the machine, which utilizes quantitative parameters effectively for short speech samples.

Key words: *Acoustic analysis of emotional speech, automatic discrimination of emotion from speech, emotion in speech, human emotion recognition*

Introduction

It is well known that emotions play a significant role in social interaction, both displaying and regulating behavior. In the phonetic sciences, the vocal expression of emotion has been researched over a long period of time, and the vocal parameters of emotions are now understood relatively well (1). Also speech scientists and engineers are now taking increasing interest in the role of the expression of emotion in voice communication; witness, for example, the ISCA Workshop on Voice and Emotion (Newcastle, Northern Ireland, 2000) and the related publications (2). The study of the vocal expression of emotion is now reaching a level of maturity where the main focus is on important applications, particularly those involving human-computer interaction in various forms.

In the area of information retrieval, the potential of prosodic/acoustic cues in signaling different affective speaker-states is gradually receiving more attention in the form of prosodic data mining—applications have been developed for major languages such as English. The automatic discrimination/classification of emotions can open up interesting new possibilities for speech corpus search engines and internet database technologies in general (note that, in this paper, the terms 'classification' and 'discrimination' are used interchangeably). There has also been some research in the area of automatic classification of emotion for minor languages: Seppänen et al. (3) demonstrated that the automatic classification of affect in spoken Finnish is possible with a reasonably high classification rate (70%–80%) for four of five basic emotions. The speech data used in the investigation by Seppänen et al. (3) was simulated emotional Finnish (produced by professional actors), and the speech segments used in the classification experiment were units containing two to four sentences with semantically neutral content. It is known that human listeners can reach even higher performance levels: listeners can discriminate between major emotions with an average accuracy of 60%–100% in utterance-length units without any lexical or syntactic cues of affect (4).

Correspondence: Dr Juhani Toivanen, MediaTeam, Department of Electrical and Information Engineering, Information Processing Laboratory, PO Box 4500, FIN-90014 University of Oulu, Finland. Fax: +358 8 553 2534. E-mail: juhani.toivanen@ee.oulu.fi

ISSN 1401-5439 print/ISSN 1651-2022 online
DOI: 10.1080/14015430500293926

Generally speaking, research on the vocal expression of emotion (from the perception/production viewpoint) has been largely based on scripted material representing utterance-length (or longer) speech units: typically, emotions are simulated during reading out an emotionally neutral text—a sentence or a short passage (4,5). Also in automatic classification experiments, the focus has been on utterance-length units (6,7). An unexplored issue is whether (and to what extent) the vocal expression of emotion can be successful in considerably shorter speech units.

The aim of this paper is to investigate the vocal expression of emotion in a unit consisting of a vowel extracted from running Finnish speech, from the viewpoint of both human perception and automatic classification. It is assumed that a vowel-length unit is potentially an effective carrier of affective content in speech, but that the perception of emotional content from such a short unit is different from that in longer discourse-level units. An interesting question is whether there are differences between human listeners and the computer in emotion classification regarding vowel-length speech stimuli.

Data

The speech data consisted of multiple repetitions of the following emotions in Finnish speech: 'neutral', 'sadness', 'joy', 'anger' and 'tenderness'. The data were produced by professional actors: nine actors (five men and four women, aged between 26 and 45 years). None of the subjects had any known pathologies of the larynx or hearing. The subjects read out a passage of some 80 words from a Finnish novel admitting several emotional interpretations; the average duration of the read passage (in the neutral state) was approximately one minute. Each speaker produced, in a random order, ten renditions of each emotional state (however, an emotional state was never repeated in the next rendition). The subjects were not given any detailed instructions concerning the emotional expressions (e.g., concerning the distinction between cold anger and hot anger). There were a total of 450 emotional speech samples (ten samples for five emotions by nine speakers). The procedure of data collection was in accordance with the Helsinki Declaration of 1975, as revised in 1983; this also holds for the listening experiments described below.

The recordings were made in an anechoic chamber using a high quality condenser microphone (Bruel & Kjaer 4188) and DAT recorder (Sony DTC-690) to obtain a 48-kHz, 16-bit recording. The microphone was placed at a fixed distance of 50 cm from subjects' lips. To compute intensity measurements, a calibration signal (provided by Bruel & Kjaer 4231) was recorded onto each tape.

The particular five emotions were chosen since it can be presumed that these emotional states represent both positive and negative valence, and high and low psycho-physiological activity levels. It can be assumed that the voice source characteristics are rather similar in certain emotions as the voice source reflects differences in the activity levels. Sadness and tenderness have a low level of activity, while joy and anger have a high level of activity. Sadness and anger have a negative valence value, while joy and tenderness are positive emotions in terms of valence. The emotions, perhaps excluding tenderness, can be assumed to represent the most important affective states, which are often called basic emotions. It is largely agreed that at least fear, anger, happiness, sadness, surprise, and disgust are among the basic emotions (8). The basic emotions are considered 'basic' because they are seen to represent survival-related patterns of responses to events in the environment, and these patterns have become universal over the course of man's evolutionary history.

Acoustic analysis

The acoustic analysis focused on a unit consisting of a vowel extracted from running Finnish speech: the first [a:] vowel from the Finnish word [ta:k:ahan] was extracted from the text using the SoundSwell software. The samples with harmonic distortion were rejected, as were samples with so weak or irregular a signal that the analyzing software could not find a period. A total of 171 samples were chosen. The analysis thus focused on [a:] from the word *taakkahan* ('indeed a burden') in the passage. The sentence context for [a:] was: *Taakkahan se vain on* ('It is indeed a burden only'). The unit of analysis, [a:], is a double vowel in a primary stressed position. As plosives are not aspirated in Finnish (i.e., they convey weak acoustic cues), and the first syllable of the word is always stressed, it can be assumed that the vowel-length unit investigated here carries reliable prosodic markers (word boundary cues, emotional cues, etc.).

Alpha ratio, standard deviation of jitter, average jitter (jitter = period to period variation in period length), shimmer (period to period variation in period amplitude), S/N (signal-to-noise) ratio and duration of the samples were measured with a signal analysis system named Intelligent Speech Analyser (ISA), developed by Raimo Toivonen, MScEng. Alpha ratio describes the spectral energy distribution (9) and thus the voice quality along the axis from hypofunctional (breathy) to hyperfunctional (strained). Alpha ratio was calculated by subtracting

the SPL (Second Pressure Level) in the range 50 Hz–1 kHz from SPL in the range 1–5 kHz.

The vowels were inverse filtered using the Iterative Adaptive Inverse Filtering (IAIF) method (10). In this method the glottal airflow is estimated from the speech pressure signal and expressed on an arbitrary amplitude scale. The glottal waveform was parameterized by calculating a robust time-domain voice source parameter NAQ (Normalized Amplitude Quotient) (11). NAQ is determined from two amplitude-domain measures, the peak-to-peak AC (Alternating Current) flow (fac) and the absolute amplitude of the negative peak of the first derivative of the flow waveform (dpeak) as follows: NAQ = fac/(dpeakT), where T denotes the fundamental period length. It has been shown that NAQ correlates strongly with the closing quotient of the glottal flow, but it can be computed in a more consistent manner because there is no need to extract the time-instant of the glottal closure (11).

Classification experiment: human listeners

The 171 samples were presented for evaluation to 50 university students and teachers (41 females, 9 males, mean age 28.5 years) in a well-damped studio using a digital recorder (Tascam DA-20) and a high-quality loudspeaker (Genelec Biamp 1019 A). The listeners were located at approximately 2.5 m distance from the loudspeaker. The samples (99 from the males and 72 from the females) were replayed at a normal conversational loudness throughout the test. The listeners' task was to recognize the expressed emotions. A forced choice questionnaire was used with the same alternatives as the expressed emotions. The test took 30 minutes altogether.

Table I shows the results of the listening test in the form of a confusion matrix: the intended emotions are shown as column indices and recognized emotions as row indices. The emotions were discriminated with an average accuracy of 37.7%.

A chance level for a five-category classification design would have been 20%; clearly, each emotion was recognized well above this level.

Table I. Results of the human emotion classification experiment: confusion matrix.

	Neutral	Sadness	Joy	Anger	Tenderness
Neutral	49.67%	15.16%	14.05%	11.43%	9.69%
Sadness	21.54%	42.91%	10.66%	8.94%	15.95%
Joy	15.67%	22.42%	28.23%	24.48%	9.21%
Anger	24.49%	21.77%	10.44%	38.17%	5.13%
Tenderness	19.25%	35.32%	12.78%	3.90%	28.75%

Classification experiment: computer

The emotional content of the selected 171 samples was classified using automatic classification methods. Software was developed for automatic classification using kNN classifier (k = 1,3,5 and the acoustic/prosodic parameters as dimensions). The k-Nearest-Neighbor classifier (kNN) is applied as a standard non-parametric method in statistical pattern recognition. The kNN is a prototype-based method, which means that a set of prototypical feature vectors from each class is stored in the classifier memory. The prototype vectors include class information. An unknown vector is then compared to all prototypes, and k closest (in vector space) prototype vectors are picked up. Majority voting is performed among these to identify the class in which most of the k closest prototypes belong. The unknown feature vector is decided to belong to that class. The classifier has one parameter, k, which is set manually by experimentation. The prototype class information is given in training based on the chosen base truth data (e.g., intended emotions during the recording in this case).

Leave-one-out is used for evaluating classifier performance. In our implementation, performance testing was included in the feature selection process in such a way that, with any feature combination, the leave-one-out testing was performed with the entire database. The average classification accuracy was used as the criterion for optimality. Table II shows

Table II. Results of the automatic emotion classification experiment (k = 1,3,5).

DIM	k	correct	feature
1	1	0.3333	alpha ratio
2	1	0.4583	duration
3	1	0.5556	shimmer
4	1	0.5417	average jitter
5	1	0.5278	SD jitter
6	1	0.4583	SN ratio
7	1	0.4861	NAQ
1	3	0.4028	duration
2	3	0.5278	alpha ratio
3	3	0.5139	SD jitter
4	3	0.5278	SN ratio
5	3	0.5417	NAQ
5	3	0.5694	average jitter-SD jitter
6	3	0.5139	shimmer
7	3	0.5000	jitter
1	5	0.3611	duration
2	5	0.5278	alpha ratio
3	5	0.5417	shimmer
4	5	0.5278	NAQ
4	5	0.5417	SN ratio-shimmer
5	5	0.5278	shimmer
6	5	0.4584	SD jitter
7	5	0.4444	average jitter

the results of the classification experiment: the rows demonstrate how the classifier performance improves as more high-performing features are added to the feature vector. The left column represents the number of features in the vector, i.e., the dimension of the vector. The second column indicates the value of k. The third column represents the average accuracy with the feature vector. In the last column on the right, listing the selected features, a prefix (-) is given with some feature names: the prefix indicates that a previously selected feature was deleted, as the backward-forward selection process increases vector dimension by one.

The automatic speaker-independent classification reached a peak level of 56.9% (k = 3), with the following five dimensions (in the order of classification power): duration, alpha ratio, signal-to-noise ratio, NAQ, and average jitter (note that the standard deviation of jitter was deleted at dimension five). The classification levels for the separate emotional states are shown in Table III.

It can be seen that all the emotions were recognized above the chance level (20%), anger having the lowest recognition level.

Discussion

An average accuracy percentage of about 60% can be obtained in experiments where listeners are to infer emotional content from vocal cues only. In a recent large-scale cross-cultural study involving European, American (US) and Asian contexts, an accuracy rate of 66% was found for such emotions as neutral, anger, fear, joy, sadness, disgust and surprise (12). In a Western cultural context (US and Europe), the vocal recognition for the basic emotions was 62% (74% for neutral, 77% for anger, 61% for fear, 57% for joy, 71% for sadness, and 31% for disgust). Scherer (13) concludes that emotion classification from voice samples produced by actors (i.e., simulated emotion portrayals) varies between 55% and 65%.

In the present investigation, the human emotion classification performance level from speech (38%) was clearly below the level suggested by Scherer (13). Of course, there is one major difference in experimental design between the present study and those reviewed by Scherer: in none of the studies referred to by Scherer is a vowel-length unit investigated in terms of the vocal expression of emotion—the existing literature focuses on considerably longer units of speech. We therefore cannot say that the test subjects 'failed' in any way in their attempt to classify emotions on the basis of the short speech stimuli. On the contrary, it is significant that the listeners were able to infer the emotional coloring well above the chance level from such short speech units. An interesting observation is that the listeners recognized the neutral emotion most accurately (50%), while sadness (43%) and anger (38%) were much more difficult; according to Scherer (13), however, sadness and anger should be easily recognizable. Indeed, in other studies, anger has been one of the easiest emotions in speech in terms of recognition, as the following classification levels suggest: 96% (14), 83% (5), and 89% (15). Also, Laukkanen et al. (16) found that, perceptually, emotional states are condensed into sadness or anger, depending on the hypo- or hyper-functionality of the voice source quality. That real-life anger should be easy to recognize (from any behavioral cues) is, of course, a very probable feature of human perceptive skills for purely evolutionary reasons. In our study, joy was the most problematic emotion in terms of recognition (28%)—this is in line with Scherer (13). Tenderness was also difficult (29%), but it must be noted that this affective state is probably outside the group of basic emotions.

It must be remembered that the emotional speech stimuli represented simulated emotional states. We cannot know how much this detracted from the power or plausibility of the emotional expressions from the listeners' viewpoint. Maybe the listeners were reluctant to recognize powerful (i.e., non-neutral) emotions from speech samples which they knew to represent simulated data? Real emotional speech material—data which are extremely difficult to come by—would be needed to decide this issue.

The automatic speaker-independent classification of emotions was very successful with an average accuracy rate of 57%. The literature on the automatic classification of affect suggests that, in a speaker-independent design, the discrimination rate for four or five emotions does not exceed 60%. McGilloway et al. (6) suggest as a 'rough benchmark' that 50% correct speaker-independent classification is a reasonable goal for automatic (speaker-independent) systems discriminating among five emotions. In a similar vein, summarizing research on the subject, ten Bosch (2) concludes that the automatic speaker-independent classification performance for three basic emotions can, at best,

Table III. Results of the automatic emotion classification experiment: confusion matrix.

	Neutral	Sadness	Joy	Anger	Tenderness
Neutral	72.22%	11.11%	16.67%	0.00%	0.00%
Sadness	43.75%	50.00%	0.00%	0.00%	6.25%
Joy	21.43%	14.29%	57.14%	7.14%	0.00%
Anger	37.50%	0.00%	25.00%	25.00%	12.50%
Tenderness	6.25%	18.75%	6.25%	6.25%	62.50%

reach a level of 70%: 'without reference to the text content of an utterance, a score of 60%–70% is about the best one can get in a speaker-independent, limited happiness/joy, anger, sadness/grief discrimination task' (p. 222). In the present investigation, the suggested 50%–60% classification rate was obtained for five emotions but, again, it must be stressed that the emotional speech samples were extremely brief in duration in comparison with the speech units commonly used in automatic classification experiments. It can be concluded, then, that very good results were achieved for the automatic classification of emotions in very short (vowel-length) units of speech.

The features used in the classification experiment turned out to be useful parameters, which the algorithm could successfully utilize in the discrimination task. It can be noted that, of the seven parameters measured and used in the classification, only five were needed to obtain the best result (57%); adding the standard deviation of jitter and shimmer to the feature vector actually lowered the classification success level. It can also be seen that a very good classification result (53%) could be obtained with only two parameters: duration and alpha ratio. Only marginal improvement was achieved with the additional three parameters. Thus, by using features of duration and alpha ratio, the speaker can effectively convey different affective states in terms of activity and valence, even in very brief speech units. It could be assumed that alpha ratio and NAQ behave in the same way as parameters, but the results suggest that alpha ratio also captures resonance effects, in contrast with NAQ. Neutral was the easiest emotional state from the viewpoint of automatic classification (72%), and joy (57%) and sadness (50%) were also relatively easy. It is plausible that duration and alpha ratio, reflecting the overall activity/passivity and effort/energy in voice, are the most effective markers of a number of basic emotions in speech.

It is interesting that the computer was clearly worse than the human listeners at recognizing angry speech samples (25% versus 38%); all the other emotions were recognized better by the computer. This may suggest that we need to modify the point we made earlier: for the human listeners, anger was, after all, a 'special emotion' from the perception viewpoint. It is probably not without some significance that the human listeners outperformed the computer with this emotion.

All in all, there was a striking difference in classification performance between human listeners and the computer (38% versus 57%). A limited number of acoustic/prosodic parameters is apparently enough for the computer, but the same does not seem to hold for human listeners. It is possible that human listeners need longer speech samples (i.e., real linguistic discourse-level contexts) for more reliable emotion discrimination than the machine, which effectively utilizes quantitative parameters even for short speech samples. That is, the computer uses brute force, which enables a decoding of emotion with a minimal speech context. Another explanation is that human listeners may really need genuine emotional speech data for a more reliable discrimination performance. With simulated data, listeners are not necessarily motivated to make analyses as keenly as they (automatically) would with real-life data. The computer, however, does not make such distinctions—hence the better performance level with the automatic classification experiment. However, the fact that the human listeners were better than the machine at recognizing anger complicates the conclusions: maybe human listeners are not motivated to analyze the emotional content of simulated speech data very thoroughly, but they are more tuned into anger than the computer, which treats all emotions on a neutral basis. Of course, genuine emotional speech data would be needed to shed light on this issue but even then the true motivation of listeners to infer the emotional content as accurately as possible would be something of a question mark—unless genuine emotional speech data is interpreted in a real situation with the possibility of withdrawal accompanying sadness, a physical threat accompanying anger, etc. (a somewhat unthinkable experimental design).

Conclusion

It has been shown that, for vowel-length units in spoken Finnish, the automatic classification of basic emotions is possible with a good accuracy rate. The classification can be carried out with a limited number of acoustic/prosodic parameters. Human listeners can recognize the emotional content well above chance level but their performance level is clearly below that of the machine. We suspect that human listeners are not necessarily disposed to recognize strong emotions in a minimal linguistic context: simulated emotional speech does not offer cues which are strong enough, or listeners do not find it necessary to be on the lookout for extreme emotions in an experimental situation. The computer, by contrast, operates without such limitations.

Acknowledgements

This research project has been financially supported by Academy of Finland (project numbers 200859, 200807, and 200997).

References

1. Murray IR, Arnott JL. Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J Acoust Soc Am. 1993;93:1097–108.
2. ten Bosch L. Emotions, speech and the ASR framework. Speech Comm. 2003;40:213–25.
3. Seppänen T, Väyrynen E, Toivanen J. Prosody-based classification of emotions in spoken Finnish. Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003, Geneva) 2003. p. 717–20.
4. Laukkanen A-M, Vilkman E, Alku P, Oksanen H. Physical variations related to stress and emotional state: a preliminary study. J Phon. 1996;24:313–35.
5. Abelin Å, Allwood J. Cross-linguistic interpretation of emotional prosody. Proceedings of the ISCA Workshop on Speech and Emotion (Belfast 2000). p. 110–13.
6. McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S. Approaching automatic recognition of emotion from voice: a rough benchmark. Proceedings of the ISCA Workshop on Speech and Emotion (Belfast) 2000. p. 207–12.
7. Yu F, Chang E, Xu Y-Q, Shum H-Y. Emotion detection from speech to enrich multimedia content. Proceedings of the 2nd IEEE Pacific Rim Conference on Multimedia (Beijing) 2001. p. 550–7.
8. Cornelius RR. The Science of Emotion. Research and Tradition in the Psychology of Emotion. New Jersey: Prentice-Hall; 1996.
9. Frøkjær-Jensen B, Prytz S. Registration of voice quality. Brüel & Kjær Techn Review. 1973;3:3–17.
10. Alku P. Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. Speech Comm. 1992;11(2–3): 109–18.
11. Alku P, Bäckström T, Vilkman E. Normalized amplitude quotient for parametrization of the glottal flow. J Acoust Soc Am. 2002;112(2):701–10.
12. Scherer KR, Banse R, Walbott HG. Emotion inferences from vocal expression correlate across languages and cultures. J Cross-Cultural Psychol. 2001;32(1):76–92.
13. Scherer KR. Vocal communication of emotion: A review of research paradigms. Speech Comm. 2003;40:227–56.
14. Dellaert F, Polzin T, Waibel A. Recognizing Emotion in Speech. Proceedings of the Fourth International Conference on Spoken Language Processing (Philadelphia 1996). p. 786–9.
15. Montero JM, Gutierrez-Arriola J, Palazuelos S, Enriquez E, Aguilera S, Pardo JM. Emotional Speech Synthesis: From Speech Database to TTS. Proceedings of the International Conference on Spoken Language Processing (Sydney 1998). p. 923–6.
16. Laukkanen A-M, Vilkman E, Alku P, Oksanen H. On the perception of emotions in speech: the role of voice quality. Logoped Phoniatr Vocol. 1997;22(4):157–68.

Logopedics Phoniatrics Vocology. 2006; 31: 153–156

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

The role of F3 in the vocal expression of emotions

TEIJA WAARAMAA[1], PAAVO ALKU[2] & ANNE-MARIA LAUKKANEN[1]

[1]*Department of Speech Communication and Voice Research, University of Tampere, Finland, and* [2]*Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, Finland*

Abstract
The present study investigates the role of F3 in the perception of valence of emotional expressions by using a vowel [a:] with different F3 values: the original, one with F3 either lowered or raised by 30% in frequency, and one with F3 removed. The vowel [a:] was extracted from the simulated emotions, inverse filtered and manipulated. The resulting 12 synthesized samples were randomized and presented to 30 listeners who evaluated the valence (positiveness/negativeness) of the expressions. The vowel with raised F3 was perceived more often as positive than the sample with original ($p = 0.063$), lowered ($p = 0.006$) or removed F3 ($p = 0.066$). F3 may affect perception of valence if the signal has sufficient energy in high frequency range.

Key words: *Emotion, inverse filtering, synthesis, third formant, valence perception*

Introduction

Most studies concerning the expression of emotions in speech have focused on the role of fundamental frequency (F0), sound pressure level (SPL), speech rate, segment duration and overall prosody (1). The role of voice quality in conveying emotions has been studied to a lesser extent. Voice quality, 'the auditory colouring of a speaker's voice' (2), can be defined as a combination of voice source characteristics (an airflow pulsation resulting from vocal fold vibration) and vocal tract (formant frequencies). According to Trojan (3), voice quality characteristics are used dualistically in responding to pleasant and unpleasant stimuli. A pleasant stimulus is related to a lax, tender voice quality while an unpleasant stimulus is prone to evoke resistance which is reflected in tense voice quality. In the production of tense voice, the larynx tends to rise, and as a consequence of a shortened vocal tract, F3 and F4 tend to increase (4). Lip spreading as in smiling also shortens the vocal tract and raises the higher formants (5).

The results of Laukkanen et al. (6) suggested that voice source characteristics (relative open time of the glottis and speed quotient) seemed to communicate the psychophysiological activity level related to an emotional state, while formant frequencies seemed to be used to code valence of the emotions, e.g., whether the emotion is positive or negative. The higher formants, F3 and F4, seemed to have greater values in positive emotions and lower in negative emotions. A recent study (7) has suggested that F3 especially is used in valence coding in [a:]. The present study investigates the perceptual role of F3 in [a:] by applying synthesis for two different voice source qualities, hypofunctional and hyperfunctional voice, with no aspiration noise.

Materials and methods

Samples

The material for the present study was derived from an earlier investigation where nine professional actors in an anechoic room recorded a prose extract expressing tenderness, sadness, anger, joy and neutral emotional states (see (8)). The first [a:] vowel was extracted from a Finnish word [ta:k:ahan], and presented to 50 listeners whose task was to identify the emotions expressed. The best recognized

Correspondence: Teija Waaramaa, MA, Department of Speech Communication and Voice Research, University of Tampere, FIN-33014 Tampere, Finland. Fax: +358 3 3551 6063. E-mail: teija.waaramaa@uta.fi

(accepted 7 November 2005)

ISSN 1401-5439 print/ISSN 1651-2022 online
DOI: 10.1080/14015430500456739

samples of sadness, tenderness and anger (three in total) were chosen for the material of the present experiment. These samples were chosen since they represent both positive and negative valence and they correspond to both hypofunctional (sadness and tenderness) as well as hyperfunctional (anger) voice qualities.

Synthesis and manipulation

The original [a:] vowels were firstly analyzed by separating the signal into the glottal flow and the vocal tract filter according to the source-filter theory (9). This separation was computed with an inverse filtering technique, the Iterative Adaptive Inverse Filtering method (IAIF) (10). The voice source was parametrized by calculating the normalized amplitude quotient (NAQ) (11). NAQ measures the relative time of the glottal closing phase from two amplitude domain values, the peak-to-peak AC flow (f_{ac}) and amplitude of the negative peak (d_{peak}) of the first derivative of the flow, and it is defined as follows: $NAQ = f_{ac}/ (d_{peak}T)$, where T is the fundamental period length. Alku et al. (11) have shown that NAQ is high in hypofunctional voice and low in hyperfunctional voice. Using the glottal flows and vocal tract models obtained in the inverse filtering stage, synthetic vowels were generated for the perceptual analysis. In order to synthesize vowels with different values of F3, the third resonance of the all-pole vocal tract model given by inverse filtering was modified by both raising and lowering its value by 30%. In addition, a vocal tract setting was computed where the F3 was completely removed. Hence, for each [a:] vowel in the three different emotions, four variants of the vocal tract settings were computed: the original vocal tract given by inverse filtering, that with F3 30% higher than in the original, that with F3 30% lower than in the original, and, finally, that with F3 completely removed. Modification of the vocal tract filter was computed by preserving the formant amplitudes of all formants. Finally, the synthetic [a:] vowels were generated by filtering the obtained glottal flows through the modified vocal tract filters in each emotion. Figure 1 shows the glottal flow waveforms in the time-domain in each emotion, and Table I shows the F3 values of the 12 resulting samples.

After the modifications, the resulting12 samples were normalized at 70 dB SPL to be presented in a listening test.

Listening test

The samples were evaluated by 30 listeners (24 females, 6 males, mean age 35 years). The Judge computer program, developed by Svante Granqvist (KTH, Stockholm) was used. The twelve [a:] samples were presented via Sennheiser HD 530 II headphones at normal conversational loudness in randomized order. Six of the samples were repeated to enable the study of the intrarater reliability. The listeners' task was to evaluate the level of positiveness or negativeness of the expressions on a visual analog scale (VAS) on an axis positive–negative (0–100 mm, the neutral point at 50 mm). The listeners had the opportunity to listen to the samples as many times as they wanted. The Judge program transferred the answers directly into numerical form (mm VAS).

Intra- and interrater reliability of the listening test was studied by calculating Cronbach's alpha (SPSS-11 software). Differences in the perception of original and modified samples were studied with Wilcoxon Signed Rank *t*-test. Relations between perceived valence and F3 were studied with the Pearson correlation.

Results and discussion

Reliability of the listening test was rather low (Cronbach's alpha 0.586 for interrater reliability and 0.560 for intrarater reliability). To some extent, this may be related to the use of short (200 ms) and synthetic samples. This makes the task demanding for the listeners. However, the results of Laukkanen et al. (6,12) suggest that it is possible to judge the valence of an emotional expression from samples as short as these. The results of Wambacq et al. (13) also show that processing of valence occurs much faster (as fast as 160 ms) than recognition of the emotion (360 ms). These results suggest that

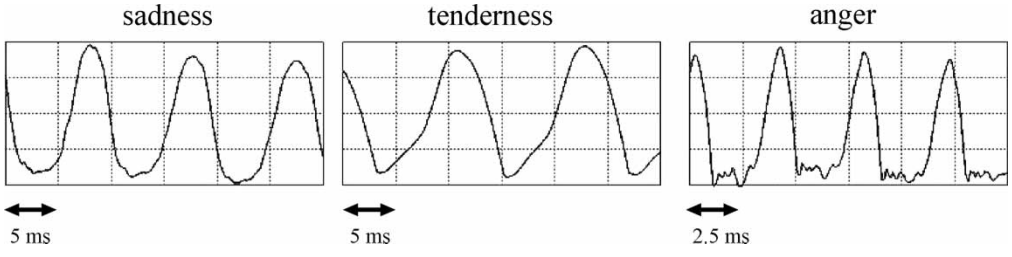


Figure 1. Glottal flow waveforms in sadness, tenderness and anger. Y-axis is flow expressed in arbitrary units.

Table I. Original and modified F3 values (in Hz) of different emotions.

	Original F3 (Hz)	F3 raised by 30% (Hz)	F3 lowered by 30% (Hz)
Tenderness	2700	3510	1890
Sadness	2630	3419	1841
Anger	2910	3783	2037

emotional characteristics in speech are processed 'non-voluntarily' (involuntarily). Figure 2 illustrates the results of the listening test.

Understandably, the results of the listening test suggest hesitation and uncertainty among the listeners in rating the samples. This can be seen in the small differences in the average valence ratings between samples and in the large standard deviation (Figure 2). Consequently, the modifications of F3 frequencies did not correlate with the responses. Furthermore, anger has been perceived as positive in valence. This may be connected to the fact that in the vocal expression of anger and happiness, the amount of high frequency energy is typically large (14), and both of them are high in intensity and pitch (15). A large amount of sound energy in the upper frequencies results from a fast flow declination rate and reflects fast glottal closing (e.g., (9,16)). As NAQ reflects the closing rate, it is plausible that NAQ is small in a hyperfunctional expression. Here NAQ was smallest for anger.

All the modifications of sadness were perceived as negative. However, the histograms illustrate that the perception of sadness with raised F3 was on average somewhat more positive than perception of the other

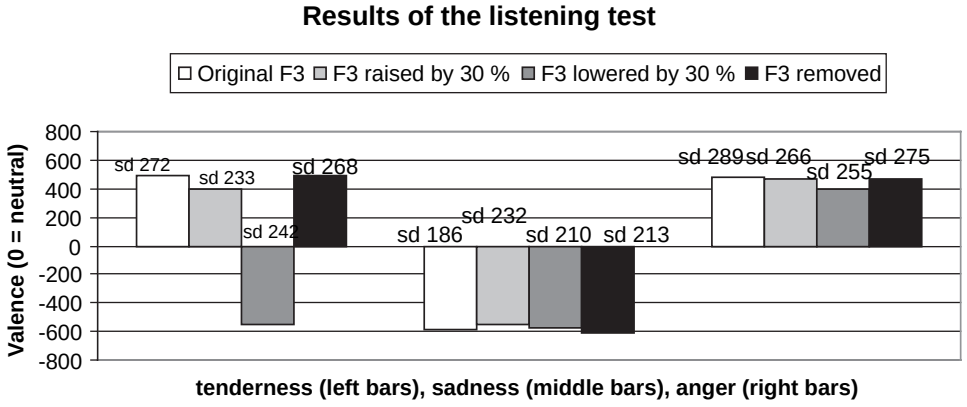


Figure 2. Perception of valence on the visual analog scale for three emotions (sd denotes standard deviation).

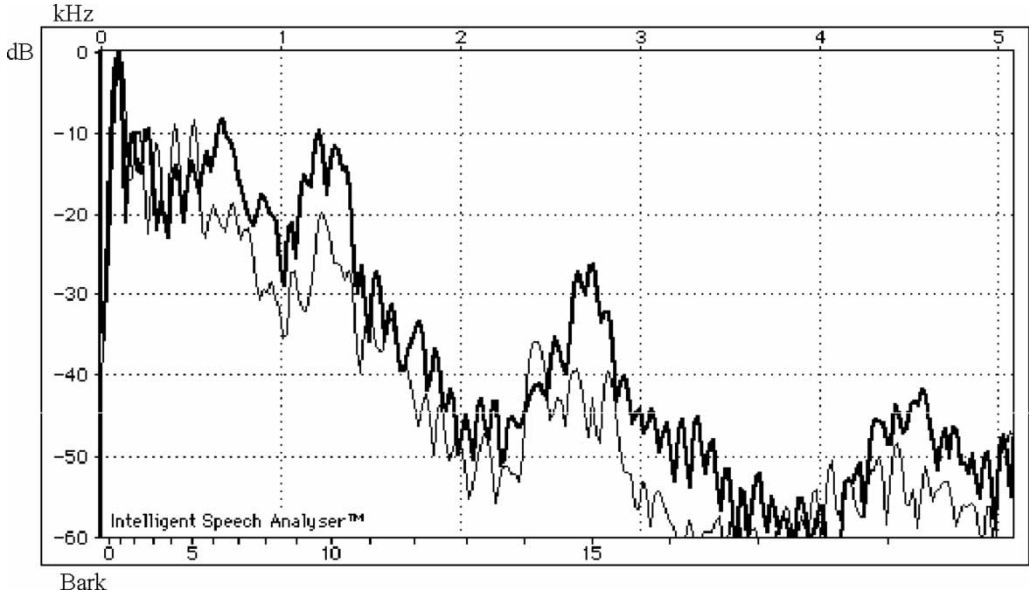


Figure 3. Long-term-average spectra of the vowel [a:] in tenderness (thick line) and in sadness (thin line), both with original values of F3. Horizontal upper axis: Frequency in kHz (horizontal lower axis: Bark scale), vertical axis: Relative amplitude in dB. The spectra were made with ISA. (ISA (Intelligent Speech Analyser) is a signal analysis system developed by Raimo Toivonen (M.Sc.Eng.)).

sadness samples, although the difference was not significant. This may be due to the fact that the voice quality in sadness was so hypofunctional and thus the spectral slope so steep that the perceptual importance of F3 in the higher frequency area seems not to be significant.

Tenderness was on average perceived as positive, except for one modification (F3 lowered by 30%). However, some significant differences were found: the samples of tenderness with raised F3 were perceived more often as positive than the samples with original ($p = 0.063$) or lowered ($p = 0.006$) or removed F3 ($p = 0.066$).

Figure 3 shows that the spectral slope of the vowel in sadness with original F3 (thin line) is somewhat steeper than the spectral slope of the vowel in tenderness with original F3 (thick line). The spectra seem to suggest that there is more energy in tenderness than in sadness, not only in the area of F3 but in the lower formant frequency area as well. This may have some importance in valence perception.

The results of this experiment would suggest that the role of F3 alone is not crucial in determining the perceived valence or they may also reflect difficulties in the perception of short synthesized samples. The role of other formant frequencies and the possible interplay with voice source characteristics and formant frequencies in the conveying of valence warrants a further study.

Acknowledgements

This study was supported by the Academy of Finland (grant nr 200807 and nr 200859).

References

1. Scherer KR. Vocal communication of emotion: A review of research paradigms. Speech Communication. 2003;40:227–56.
2. Laver J. The phonetic description of voice quality. Great Britain: Cambridge University Press; 1980.
3. Trojan F. Experimentelle Untersuchungen über den Zusammenhang zwischen dem Ausdruck der Sprechstimme und dem vegetativen Nervensystem. Folia Phoniatr (Basel). 1952; 4(2):64–92.
4. Sundberg J, Nordström P-E. Raised and lowered larynx—the effect on vowel formant frequencies. Quarterly progress and status report. Department of Speech, Music and Hearing, The Royal Institute of Technology, Stockholm; 1976. p 35–9.
5. Titze IR. Principles of voice production. Second printing. Iowa City, USA: National Center for Voice and Speech; 2000.
6. Laukkanen A-M, Vilkman E, Alku P, Oksanen H. On the perception of emotions in speech: the role of voice quality. Scandinavian Journal of Logopedics, Phoniatrics, Vocology. 1997;22:157–68.
7. Waaramaa T, Laukkanen A-M, Alku P, Björkner E, Leino T. Perception of emotions in mono-pitched vowels. Unpublished observation.
8. Airas M, Alku P. Emotions in short vowel segments: Effects of the glottal flow as reflected by the normalized amplitude quotient. Phonetica. In press.
9. Fant G. Acoustic theory of speech production. With calculations based on X-ray studies of Russian articulations. 2nd ed. The Hague: Mouton; 1970.
10. Alku P. Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. Speech Communication. 1992;11:2–3, 109–18.
11. Alku P, Bäckström T, Vilkman E. Normalized amplitude quotient for parametrization of the glottal flow. J Acoust Soc Am. 2002;112(2):701–10.
12. Laukkanen A-M, Vilkman E, Alku P, Oksanen H. On the perception of emotional content in speech. In: Elenius K, Branderud P, editors. Proceedings of the XIIIth International Congress of Phonetic Sciences. August 13–19. Stockholm: Department of Speech Communication and Music Acoustics, Royal Institute of Technology and the Department of Linguistics, Stockholm University 1995; 1/4, 246–9.
13. Wambacq IJA, Shea-Miller KJ, Abubakr A. Non-voluntary and voluntary processing of emotional prosody: an event-related potentials study. Neuroreport. 2004;15(3):555–9.
14. Laukka P. Vocal expression of emotion. Discrete-emotions and dimensional accounts. Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 141. Uppsala, Sweden; 2004.
15. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J Acoust Soc Am. 1993;93(2):1097–108.
16. Gauffin J, Sundberg J. Spectral correlates of glottal voice source waveform characteristics. J Speech Hear Res. 1989;32: 556–65.

Original Paper

Folia Phoniatrica et Logopaedica

Folia Phoniatr Logop 2008;60:249–255
DOI: 10.1159/000151762

Published online: September 2, 2008

Monopitched Expression of Emotions in Different Vowels

Teija Waaramaa[a] Anne-Maria Laukkanen[a] Paavo Alku[b] Eero Väyrynen[c]

[a]Department of Speech Communication and Voice Research, University of Tampere, Tampere,
[b]Department of Signal Processing and Acoustics, Helsinki University of Technology, Espoo, and
[c]MediaTeam, University of Oulu, Oulu, Finland

Key Words
Voice quality · Inverse filtering · Voice source · Formants · Perception of emotions

Abstract
Fundamental frequency (F_0) and intensity are known to be important variables in the communication of emotions in speech. In singing, however, pitch is predetermined and yet the voice should convey emotions. Hence, other vocal parameters are needed to express emotions. This study investigated the role of voice source characteristics and formant frequencies in the communication of emotions in monopitched vowel samples [a:], [i:] and [u:]. Student actors (5 males, 8 females) produced the emotional samples simulating joy, tenderness, sadness, anger and a neutral emotional state. Equivalent sound level (L_{eq}), alpha ratio [SPL (1–5 kHz) – SPL (50 Hz–1 kHz)] and formant frequencies F1–F4 were measured. The [a:] samples were inverse filtered and the estimated glottal flows were parameterized with the normalized amplitude quotient [NAQ = $f_{AC}/(d_{peak}T)$]. Interrelations of acoustic variables were studied by ANCOVA, considering the valence and psychophysiological activity of the expressions. Forty participants listened to the randomized samples (n = 210) for identification of the emotions. The capacity of monopitched vowels for conveying emotions differed. L_{eq} and NAQ differentiated activity levels. NAQ also varied independently of L_{eq}. In [a:], filter (formant frequencies F1–F4) was related to valence. The interplay between voice source and F1–F4 warrants a synthesis study.

Introduction

Fundamental frequency (F_0), the main correlate of pitch, its variations and sound pressure level (SPL), which is mainly heard as loudness, are well known to be among the most important variables in emotion expressions [1, 2]. Both of them tend to increase in accordance with psychophysiological activity level, being higher in emotions with high arousal (e.g. joy and anger) and lower in emotions with less arousal (e.g. tenderness and depressive sadness) [1, 3]. In the literature on emotions it is widely agreed that there are four basic emotions (joy, anger, fear and sadness) which are universal, not culture-related. These basic or primary emotions represent both high arousal activity level (joy, anger and fear) and low arousal activity level (sadness). Additionally, there are also so-called secondary or social emotions, which are culturally related and therefore more difficult to define (such as longing, boredom or satisfaction).

Voice quality has been considered a crucial variable in differentiating emotions which are communicated

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1021–7762/08/0605–0249$24.50/0

Accessible online at:
www.karger.com/fpl

Teija Waaramaa, MA
Department of Speech Communication and Voice Research
University of Tampere
FI–33014 Tampere (Finland)
Tel. +358 3 3551 7104, Fax +358 3 3551 6063, E-Mail teija.waaramaa@uta.fi

through a subtle coloring of the voice [2, 4, 5]. Thus, voice quality is a paralinguistic means of signaling differences in meaning in speech [6]. Voice quality and F_0 may vary individually in conveying vocal emotional content [7].

Voice quality is defined by Laver [8] in a broad sense as the individual coloring of the speaker's voice, which is determined both by the voice source (a result of vocal fold vibration) and vocal tract characteristics [9]. Thus, voice quality results from both phonatory and articulatory characteristics. In order to study the voice source, inverse filtering is needed to separate the resonances from the signal. The voice source varies together with F_0, SPL (or equivalent sound level, L_{eq}) and also with tempo [10–14], and therefore, the individual role of voice quality in emotional expression has not been simple to study. However, it is reasonable to assume that the voice source may also vary independently, not merely related to variation in F_0 and SPL [15, 16]. Moreover, F_0 and SPL seem to interact, SPL typically rising together with F_0. F_0 is also one of the means to increase SPL [17].

In singing, the variations in pitch are always predetermined, and the relative loudness and duration are also more or less predirected. Yet the vocalist's voice should be emotionally expressive. Thus, singers, compared to speakers, need to use different strategies to convey emotional expressions to the audience. This acoustic conveyance of emotions has been investigated among other things by analyzing the vocalist's interpretation *of the music* and the expression of the emotions *of the character* portrayed, and also by investigating the *vocalist's own psychophysiological state* while performing [18]. From the perceptual viewpoint, the listeners' individual abilities to perceive emotional quality and the acoustic characteristics of the voice signal used in the perception process are of importance [18]. Vocalists' and also actors' expressions and performance therefore need to be tightly controlled. One means of improving this control is to practice with monopitched sounds. Hence, the control concentrates on the voice quality itself, not only on the prosodic features.

As every vowel has its own vocal tract setting, it is plausible that the expression-related acoustic changes are different when they occur in the context of a different vowel. The same articulatory movement, for example moving the tongue forward in the oral cavity, causes different acoustic changes in different vowels [19].

The present study investigated voice quality parameters (voice source and formant frequencies) in monopitched emotional expressions. It may be assumed that, when the variation of F_0 is eliminated, the role of voice quality in the expression of emotions would become clearer. Different vowels were studied since it was hypothesized that there might be (1) differences between vowels in conveying emotional content and (2) differences in the possible expression-related changes in the formant structure of different vowels.

Materials and Methods

Subjects and Recordings

The material for the present study was recorded in the University of Tampere in a well-damped studio using a digital recorder Tascam DA-20 and a Brüel & Kjær 4165 microphone, placed at distance of 40 cm from the subject's lips. Thirteen graduating professional actors (5 males and 8 females) with normal voices and without any known pathologies of the larynx or hearing served as subjects. They produced in random order three monopitched prolonged steady vowels, [a:], [i:] and [u:], separated from each other, expressing randomly four emotional states: anger, joy, sadness, tenderness, and a neutral emotional state at a comfortable speaking pitch. These emotional states were chosen since they represent both high and low activity level and positive and negative emotional valences. Intensity and duration varied freely in the expressions, but vowel pitch was standardized. The material contained 195 vowel samples (13 actors × 5 emotional states × 3 vowels).

Perceptual Analysis

The samples were replayed to 40 listeners, consisting of university teachers and students (20 males, 20 females, mean age 38 years in females and 39 in males). An equal number of subjects from both genders were used to investigate possible gender differences in the perception of emotional expressions. The computer program Judge (developed by Svante Granqvist, KTH, Stockholm) was used to evaluate the samples. The participants listened to the samples using Sennheiser HD 530 II headphones. The Judge program replayed the samples in a different randomized order for every listener. All of the recorded samples (n = 195) were used in the listening test. Fifteen of them were repeated in order to study intrarater reliability. The listeners' task was to state for each of the 210 samples which emotion they perceived in the sample. A visual analog scale (0–1,000 units) was used. One end, 0, was labeled 'neutral' (no emotion), and the other end (1,000 units) was labeled according to each emotional state that was simulated in the study. The listeners were allowed to listen to the samples as many times as they felt they needed. However, the participants were recommended to listen to each sample only once, if possible, because it was of interest to study the very first and thus the basic reaction to or perception of the signal heard and hence to avoid any speculations which were prone to arise from the samples (average 2,336 ms in males and 2,472 ms in females). A confusion matrix of the listeners' answers was calculated in order to investigate the percentage of similar answers between the participants. Intrarater reliability was studied by calculating the percentage of similar answers given by the listeners to the repeated samples (n = 15). The differences between the genders in the results of the listening test were also studied.

Table 1. Correct recognition of the emotions expressed in different vowels

Emotion	[a:]	[i:]	[u:]
Neutral	46%	39%	40%
Sadness	42%	57%	58%
Joy	41%	43%	28%
Anger	73%	64%	67%
Tenderness	61%	48%	35%

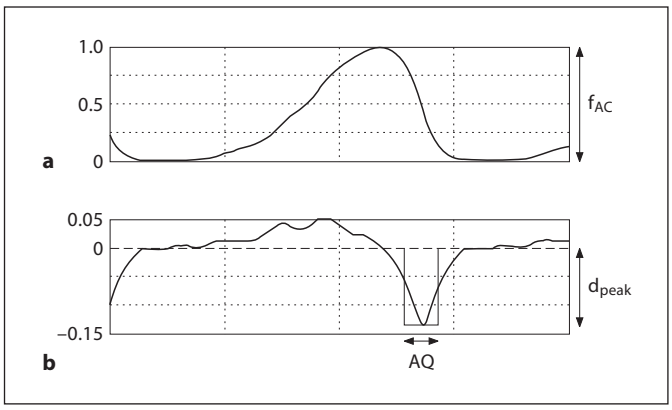


Fig. 1. Inverse filtered signal. **a** Glottal flow. f_{AC} = AC flow amplitude. **b** First derivative of glottal flow. d_{peak} = Negative peak amplitude of the derivative. AQ (amplitude quotient): f_{AC}/d_{peak}. Time on the horizontal axis, flow on an arbitrary scale on the vertical axis.

Acoustic Analysis

The 195 samples were analyzed for F_0, L_{eq}, alpha ratio and formant frequencies F1–F4 with a signal analysis system named Intelligent Speech Analyser (ISA), developed by Raimo Toivonen, MScEng. F_0 was measured to ensure that the expressions were monopitched since the actual Hz value is typically connected with the psychophysiological activity level and hence may vary along with intensity. The alpha ratio reflects voice quality by showing the sound level difference between the range above and below 1 kHz [20]. It was calculated here by subtracting the L_{eq} in the range 50 Hz–1 kHz from the L_{eq} in the range 1–5 kHz. The alpha ratio naturally depends on both the voice source and filter characteristics. Formant frequencies were measured on spectrograms and the FFT average spectra were taken from the middle portion of each vowel sample.

In order to study the voice source and formant frequencies separately, the voice signal was inverse filtered using the Iterative Adaptive Inverse Filtering method [21], which uses the acoustic speech pressure signal as the input, thereby enabling the study of natural speech without the inconvenience and restrictions inherent in using a flow mask over the face. Since inverse filtering techniques require vowels with high F1 in order to be accurate, only [a:] vowels were inverse filtered in the present study. The resulting voice source was parameterized by calculating the normalized amplitude quotient [NAQ = $f_{AC}/(d_{peak}T)$] [22]. NAQ measures the relative time of the glottal closing phase from two amplitude domain values, peak-to-peak AC flow (f_{AC}) and the amplitude of the negative peak (d_{peak}) of the first derivative of the flow. T is the fundamental period length (fig. 1). NAQ has been shown to reflect phonation type, being low in hyperfunctional (pressed) voice and high in hypofunctional (breathy) voice [22]. NAQ also correlates with SPL [22].

Statistical Analysis

Analysis of covariance (ANCOVA, SPSS-15, Chicago, Ill., USA) was used to study the interrelations of acoustic variables in emotional expressions. Two main characteristics of expressions were considered: valence and psychophysiological activity level. Valence (the affective value of an emotion on an axis positive – neutral – negative) and psychophysiological activity level (on an axis high – medium – low) were assigned arbitrary numbers by the authors to enable statistical analysis. The positive emotions joy and tenderness were assigned a positive value 1, a neutral emotional state 0, and the negative emotions anger and sadness –1. For the psychophysiological activity level joy and anger were given a positive value 1, neutrality 0, and tenderness and sadness a negative value –1. Thus, there were fewer choices (n = 3) in the study of valence and the psychophysiological activity level than in the actual emotions (n = 5).

Dependent variables were: (1) Filter (= weighted sum variable of formant frequencies F1–F4 constructed with principal component analysis), (2) NAQ (measured only in vowel [a:]), (3) L_{eq}, and (4) alpha ratio. The effect of gender was included in the models. L_{eq} was set as a covariate in order to study the dependence of other characteristics on it. Bonferroni correction was used. Of the emotions expressed in [a:], the data for 1 female subject had to be excluded, those for 4 females in [u:], and all samples expressing tenderness in vowel [u:] produced by female subjects were completely excluded due to analysis problems.

Results

Listening Test

Emotions were recognized with 50% accuracy in the listening test (table 1). The percentage for intrarater reliability was 59%. The best recognized emotion was anger in all vowels with 68% accuracy, while joy was the least recognized emotion with 37% accuracy. Sadness was the most often chosen emotion for an answer (24% of all answers), while joy was the most seldom chosen emotion (15%). There were differences in the perception between the vowels. Vowel [a:] conveyed best tenderness, vowels [i:] and [u:] sadness. Anger was well conveyed by all the vowels studied, however, vowel [a:] was a somewhat better conveyor of anger than the other vowels. Emotions conveyed by vowel [u:] were quite poorly recognized, especially positive emotions.

Table 2. Averages of L_{eq}, alpha ratio and formant frequencies F1–F4 in vowels [a:], [i:] and [u:], and NAQ averages for vowel [a:] for males and females separately

	[a:]					[i:]					[u:]				
	neutral	sadness	joy	anger	tender-ness	neutral	sadness	joy	anger	tender-ness	neutral	sadness	joy	anger	tender-ness
Males															
L_{eq}	65	58	67	72	57	62	56	64	68	56	62	59	64	68	57
Alpha ratio	−10	−10	−8	−7	−9	−14	−18	−13	−6	−18	−27	−28	−25	−23	−29
NAQ	0.11	0.16	0.1	0.08	0.18										
F1	611	603	646	650	659	306	284	461	301	323	306	314	327	319	327
F2	1,068	1,060	1,124	1,073	1,107	2,024	2,033	2,119	2,050	2,080	646	629	685	672	633
F3	2,575	2,627	2,700	2,713	2,657	2,588	2,713	2,735	2,903	2,683	2,433	2,459	2,403	2,494	2,567
F4	3,372	3,381	3,407	3,372	3,492	3,329	3,372	3,390	3,643	3,669	2,993	3,191	3,118	3,415	3,518
Females															
L_{eq}	62	55	63	72	55	60	56	56	63	66	60	56	56	70	70
Alpha ratio	−4	−6	−4	−1	−7	−15	−23	−15	−19	−12	−22	−30	−21	−26	−26
NAQ	0.11	0.15	0.14	0.1	0.17										
F1	678	692	719	778	668	377	380	382	428	425	377	393	396	458	458
F2	1,190	1,222	1,295	1,225	1,284	2,538	2,455	2,557	2,199	2,592	651	668	657	751	773
F3	2,891	3,004	3,182	2,931	3,082	3,163	3,192	3,112	3,152	3,308	2,739	2,455	2,830	2,552	2,764
F4	3,954	3,892	4,180	3,941	4,162	4,078	4,121	4,110	3,898	4,038	3,708	3,898	4,212	3,583	3,583

Valence was perceived with 70.5% and activity with 76.5% accuracy of the answers given. This result may suggest fewer difficulties in the perception of valence and psychophysiological activity level compared to the recognition of actual emotions. Valence was perceived correctly in 76% of [a:] and [i:] vowels and in 60% of vowel [u:]. Psychophysiological activity level was recognized best in vowel [i:] with 86% correct answers. The corresponding percentage for [a:] was 73% and for [u:] 71%.

Some minor nonsignificant gender differences were seen in the answers: males perceived the emotions with 48% accuracy, females with 52% accuracy on average for all emotions expressed. No gender differences in perception were found regarding emotions expressed by the same or opposite gender.

Acoustic Analysis

The averages of the parameters measured are given in table 2 separately for both genders.

Effects of Filter

In all three vowels filter (F1–F4) was related to gender, which was to be expected. In vowel [a:] filter was also related to valence ($F_{1.51} = 6.18$, $p = 0.016$), differentiating between positive and negative emotions (Bonferroni adjusted $p = 0.016$). The interaction effect between valence and gender was not significant. Valence was thus related to formant frequencies in a similar way in both genders. In positive emotions, the average frequencies of the formants tended to be somewhat higher than in negative emotions.

In vowels [i:] and [u:], unlike for [a:], filter was not significantly related to valence. As expected, formant patterns in emotional expressions differed between vowels. The perceptual effect of the formant changes was most likely related to an interplay between voice source and frequency relations between adjacent formants.

Effects of Voice Source

In all three vowels, L_{eq} was associated with activity, which was to be expected. Significant differences between low and medium (Bonferroni test, $p = 0.001$) and low and high ($p < 0.001$) activity levels were observed. In [i:] and [u:], only L_{eq} and activity level were significantly related. L_{eq} was highest in anger and lowest in tenderness and sadness in both genders. There was a high negative correlation between L_{eq} and NAQ in both males and females. NAQ was smallest in anger in both genders and greatest in tenderness. In [a:], NAQ was related to the psychophysiological activity level ($F_{1,57} = 27.9$, $p < 0.001$). Differences in NAQ also remained significant when L_{eq} was set as a covariate, suggesting differences in phonation type independent of voice loudness (fig. 2).

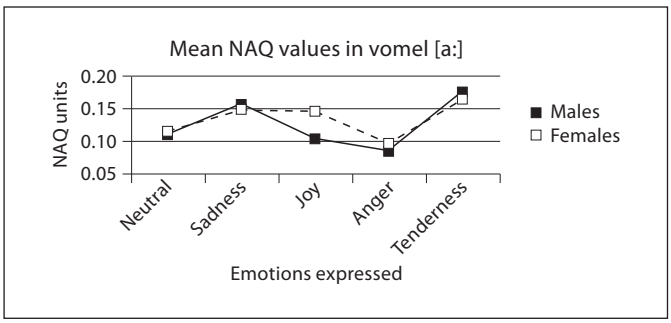


Fig. 2. Mean NAQ values in monopitched emotional expressions for males and females in vowel [a:].

Unlike L_{eq} and NAQ, L_{eq} and alpha ratio did not show any correlation. This is most likely due to the fact that alpha ratio – spectral energy distribution – is also affected by resonances.

Discussion

The perception of emotion samples revealed that the listeners were more likely to hear the negative coloring of voices than the positive coloring. This tendency may have been inherited in the course of evolution: on the one hand, humans have had to be aware of a possible threat and be sensitively attuned to negative signals especially. On the other hand, positive stimuli did not require any necessary reaction for survival. The listening test also revealed that emotion perception in different vowels varies. The open back vowel [a:] conveyed better anger, tenderness and neutrality than the other two vowels, most likely due to its evenly spread (nondiffuse) formant structure, which gives more freedom for expression. Alpha ratio was significantly higher in vowel [a:] than in [i:] and [u:], which was to be expected. This may explain the better recognition of emotions in [a:]. In all emotions studied, alpha ratio values were significantly higher in [a:] than in the other vowels. This may be due to louder formant amplitudes and greater amount of spectral energy, which may have had some perceptual relevance. Anger was conveyed remarkably well by all vowels studied, implying that it may not be related to filter functions but rather to the voice source, and may thus not be vowel-dependent. Joy was slightly better recognized in the front vowel [i:] than in [a:], but distinctively better in [i:] than in [u:]. This result for joy may be connected to the formant structure of [i:] where F2–F4 are all relatively high in frequency, which, in turn, may give a bright sound to the vowel when signaling positive emotions.

In the writing of song lyrics, the words tend to be chosen such that the vowels support the mood of the message of the song, e.g. /a/ and /e/ may be used when brightness is needed in the expressions, and /u/ and /y/ when darker voice timbre is desired. The resonances create the acoustic structures of the vowels, and so the formant frequencies differ between them. Thus, their ability to modify perceptually relevant resonances is different.

In earlier investigations, formant frequencies (F2–F4) have been found to be higher in positive valence than in negative valence [16, 23]. The higher frequencies may be due to the smiling position of the lips, which shortens the vocal tract. This effect should affect the highest formants of [i:] especially. However, sadness was signaled well by both [i:] and [u:] vowels. Their diffuse formant structures and smaller amount of spectral energy in the higher formant frequency area (alpha ratio was significantly lower than in [a:]) may account for this result. Since there was no significant difference between valence and formant frequencies in [i:] and [u:], the use of the formant amplitudes by varying L_{eq} may play a role in darkening the timbre of the voice for signaling negative emotions in vowels [i:] and [u:]. The perceptual value of a formant, its loudness, is influenced by the amplitude of the formant and the sensitivity of hearing at the frequency range of the formant. Formant amplitude, in turn, is affected by tilting of the voice source spectrum, by formant tuning (how close a match there is between a voice source partial and a formant) and by the distance of formants from each other. Formant amplitude is higher when the voice source spectrum has stronger overtones (louder voice or more pressed phonation type), and when the formant frequency decreases and hence the formant coincides with a lower and thus a stronger overtone. Furthermore, distance between formant frequencies has an effect on formant amplitudes; if two formants are close to each other, their amplitudes become 6 dB stronger and the area between them 12 dB [9].

F_0 and SPL interact with the voice source characteristics (glottal flow velocity waveform) [24, 25]. In the present study, loudness was allowed to vary, since a strict control of SPL would have affected the phonation type too much. Phonation type along the axis hypofunctional-hyperfunctional (pressed) is reflected in NAQ and spectral energy distribution, i.e. in this case in alpha ratio. Smooth, almost sinusoidal glottal flow velocity waveform (and thus high NAQ) characterizes a soft, breathy, hypofunctional phonation type, while a steeper waveform is seen

in a pressed, hyperfunctional phonation type (and therefore NAQ would be low) [9]. A smooth voice source then has a steeper spectrum slope (resulting in lower alpha ratio). If psychophysiological activity level is high, phonation type is hyperfunctional and consequently, if it is low, phonation type is hypofunctional. In emotional expressions with high psychophysiological activity level, the spectrum of the voice signal tends to be flatter and the glottal volume velocity waveform sharper than in the emotional expressions with low psychophysiological activity level [23, see also ref. 17]. In the latter, the spectrum is more tilting and the waveform is smoother, sometimes almost sinusoidal.

The grade of hypo-/hyperfunctionality of the phonation type is reflected in alpha ratio and NAQ. In the present study, voice quality was more hyperfunctional in joy and anger, and more hypofunctional in tenderness and sadness, alpha ratio being lower in tenderness and sadness and NAQ being smaller in joy and anger. Thus, the spectral slope was flatter in a more hyperfunctional and steeper in a more hypofunctional phonation type. The regulation of the vocal apparatus can be presumed to be holistic, especially in emotional expressions. In emotional expressions (here in vowel [a:]) NAQ seemed to have an independent effect, not combined with L_{eq}. This result was in line with earlier findings [3, 15]. However, no gender differences were found in NAQ, which contrasts with earlier findings [26]. Gender differences in NAQ may reflect differences in F_0 *control* mechanisms (not differences in F_0 per se, since NAQ by definition is normalized according to F_0).

Conclusions

(1) Monopitched vowels [a:], [i:] and [u:] differed in their capacity to convey emotions.

(2) In [a:], filter (formant frequencies F1–F4) was related to valence.

(3) Voice source characteristics (reflected in NAQ) appeared to have a role in expression, not merely related to L_{eq}.

(4) In vowels [i:] and [u:] L_{eq} was the only statistically significant variable in emotional expressions. Thus, either voice source and filter characteristics are used differently in different vowels or due to differences in the vocal tract setting in different vowels the same phonatory or articulatory characteristics have different acoustic consequences. It should be noted, however, that the data sample was smaller for [u:] in females.

(5) The perceptual effects of the interplay between voice source and formant frequencies in different vowels warrant synthetic study.

Acknowledgments

This study was supported by the Academy of Finland (grants No. 200807, No. 200859 and No. 200997). The authors would like to thank Hanna-Mari Pasanen, MSc, of the Unit for Science, Technology and Innovation Studies (TaSTI), University of Tampere, for statistical analysis.

References

1 Lieberman P, Michaels SB: Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. J Acoust Soc Am 1962;34:922–927.
2 Murray IR, Arnott JL: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J Acoust Soc Am 1993;93:1097–1108.
3 Laukkanen AM, Vilkman E, Alku P, Oksanen H: On the perception of emotions in speech: the role of voice quality. Scand J Log Phon Voc 1997;22:157–168.
4 Scherer KR: Vocal communication of emotion: a review of research paradigms. Speech Commun 2003;40:227–256.
5 Gobl C, Ní Chasaide A: The role of voice quality in communicating emotion, mood and attitude. Speech Commun 2003;40:189–212.
6 Campbell N, Mokhtari P: Voice quality: the 4th prosodic dimension. Proc 15th Int Congr Phonet Sci, Barcelona, August 2003, pp 2417–2420.
7 Ladd DR, Silverman KEA, Tolkmitt F, Bergmann G, Scherer KR: Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. J Acoust Soc Am 1985;78:435–444.
8 Laver J: The Phonetic Description of Voice Quality. Cambridge, Cambridge University Press, 1980.
9 Fant G: Acoustic Theory of Speech Production. With Calculations Based on X-Ray Studies of Russian Articulations, ed 2. The Hague, Mouton, 1970.
10 Granström B, Nord L: Ways of exploring speaker characteristics and speaking styles. Proc 12th Int Congr Phonet Sci, Aix-en-Provence, August 1991, pp 278–281.
11 Laukkanen AM: On Speaking Voice Exercises: a Study on the Acoustic and Physiological Effects of Speaking Voice Exercises Applying Manipulation of the Acoustic-Aerodynamic State of the Supraglottic Space and Artificially Modified Auditory Feedback; doct diss Medical School University of Tampere, Tampere, 1995.
12 Alku P, Vilkman E, Laukkanen AM: Estimation of amplitude features of the glottal flow by inverse filtering speech pressure signals. Speech Commun 1998;24:123–132.

13 Sonesson B: On the anatomy and vibratory pattern of the human vocal folds: with special reference to a photo-electrical method for studying the vibratory movements. Acta Otolaryngol Suppl 1960;156:1–80.
14 Sundberg J, Andersson M, Hultqvist C: Effects of subglottal pressure variation on professional baritone singers' voice sources. J Acoust Soc Am 1999;105:1965–1971.
15 Laukkanen AM, Vilkman E, Alku P, Oksanen H: Physical variations related to stress and emotional state: a preliminary study. J Phonet 1996;24:313–335.
16 Waaramaa T, Laukkanen AM, Alku P, Björkner E, Leino T: Perception of emotions in mono-pitched vowels; in Rantala L (ed): Puheopin laitos. Report 5. Department of Speech Communication and Voice Research, University of Tampere, 2007.
17 Alku P, Vintturi J, Vilkman E: The effect of fundamental frequency per se on vocal intensity in soft, normal and loud phonation. Proc 25th World Congr Int Assoc Logop and Phoniatr, Montréal, 2001.
18 Scherer KR: Expression of emotion in voice and music. J Voice 1995;9:235–248.
19 Fant G: Speech Sounds and Features. Cambridge, Massachusetts Institute of Technology, Colonial Press, 1973.
20 Frøkjær-Jensen B, Prytz S: Registration of voice quality. Brüel Kjær Tech Rev 1973;3:3–17.
21 Alku P: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun 1992;11:109–118.
22 Alku P, Bäckström T, Vilkman E: Normalized amplitude quotient for parametrization of the glottal flow. J Acoust Soc Am 2002;112:701–710.
23 Waaramaa T, Alku P, Laukkanen AM: The role of F3 in the vocal expression of emotions. Logoped Phoniatr Vocol 2006;31:153–156.
24 Gauffin J, Sundberg J: Spectral correlates of glottal voice source waveform characteristics. J Speech Hear Res 1989;32:556–565.
25 Sundberg J, Gauffin J: Waveform and spectrum of the glottal voice source. STL-QPSR 1978;2–3:35–50.
26 Airas M, Alku P: Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. Phonetica 2006;63:26–46.