Creation and utilisation of the MediaTeam Emotional Speech Corpus

Juhani Toivanen, Tapio Seppänen, Eero Väyrynen
MediaTeam, University of Oulu, Finland

Abstract
The *MediaTeam Emotional Speech Corpus* is currently the largest database of emotional speech for colloquial modern Finnish, containing simulated emotional content. The specific aim of the research is to investigate in detail the phonetic and phonological/linguistic correlates of basic or primary emotions in spoken Finnish, to develop statistical classification methods of emotional speech signals and to develop methods to utilise this knowledge in speech corpus search engines. The vocal expression of emotion in speech is approached utilising both the abstract phonological information on the type of pitch contour and the distribution of focal sentence accents and the concrete phonetic information on the actual implementation of F0 curves and other global (continuously variable) prosodic parameters. The annotation procedure reflects these two levels, the aim being to investigate both the *linguistic* and *paralinguistic* ways of the vocal expression of emotion. The algorithms developed for prosodic parameter estimation, the statistical classification of emotional speech, and the estimation of the degree of emotions can open up new opportunities for future Internet database technologies and methods for content-based information retrieval from audio databases.

1. Introduction

Human social communication importantly rests on exchanges of non-verbal signals, including the (non-lexical) expression of emotion through speech. It is well known that emotions play a significant role in social interaction, both displaying and regulating patterns of behaviour and maintaining the homeostatic balance in the organism. In everyday communication, certain emotional states (e.g. boredom and fear/nervousness) are probably expressed mainly only non-verbally since social conventions demand that, as a rule, negative emotions be concealed. Today, the significance of emotions is acknowledged across scientific disciplines – "Descartes' error" is thus being rectified, and emotions and feelings are no longer regarded as "intruders in the bastion of reason" (Damasio 1994).

The vocal emotion literature is large but somewhat fragmented. There exists nothing like a standard model of the expression of emotion in speech although various aspects of affect-related vocal features have been described. However, the research has concentrated on the vocal correlates of emotions in spoken *English* and some other major languages (e.g. French and German): practically nothing is known about the specific acoustic correlates of emotions in (continuous) spoken *Finnish*. So far, the expression of emotion in spoken Finnish has been studied mainly in the context of very short nonce syllables.

The investigation outlined in this paper represents basic research, and fills an obvious lacuna in the literature on the topic. On an applied side, to develop automatic methods for determining the emotional state of the (Finnish) speaker is an equally important goal, with immediate and immense application potential e.g. in situations where speech is the primary mode of interaction with the machine. Human-computer interfaces could be made to respond differently, depending on the emotional state of the user. Content-based information retrieval is a very important field of application: search engines that could locate *happy* or *sad* speech in an audio database would be very useful. One of the aims of the research described in this paper is to develop such methods.

One cannot directly extrapolate from the vocal emotion literature on other languages. The intonational structure of Finnish has not been described in detail and very little is currently known about the correlation between e.g. pitch patterns (sentence melody) and emotions in spoken Finnish. It should be stressed that, from the viewpoint of intonation, Finnish is substantially different from English, for example; one important difference is that rising tones are uncommon in non-dialectal spoken Finnish. Thus, the findings concerning the correlation between different nuclear tones and perceived emotions/attitudes reported for English in the large literature on intonation cannot, as such, be applied to Finnish.

2. Some IT background: information retrieval from digital databases

The current stages of the digital revolution – in particular, the Internet and the new ways of storing and retrieving information – have produced very large collections of digital data: for example, text, audio and video libraries. Such digital collections are now an important part of a properly functioning society.

Consequently, specific methods for effectively browsing and searching such digital libraries are badly needed.

From the viewpoint of the end user, the problem is how to find the required information from a plethora of increasingly large digital databases. The problem concerns mainly the currently booming media types, such as digital speech, music and image, where the search criteria often include semantic concepts. The current demand for automatic interpretation of digital data has resulted in intensive research on content-based retrieval. Theoretically, the central problem is how to narrow down the semantic gap between the concept-based and the content-based approaches to data indexing. The aim is thus to close the gap between the natural semantic concepts used by a person seeking information and the computer-interpretable database descriptors automatically derived from the data contents.

Information retrieval applications are based on an understanding of the content of the file(s) and there are, basically, two approaches. On the one hand, human interpretation can be utilised to generate semantic keywords but this tends to be inconsistent over time, and is very expensive. On the other hand, automatic interpretation of data can be developed: it is cheap, replicable and scalable. The automatic interpretation may be wrong but it is consistently so: the search results, whether they be good or bad, are at least systematic.

As for the development of search engines, a critical fact is that the digital databases are so massive: the organising and indexing of the material is very demanding, and can never be done manually. Therefore, at least semiautomatic means for the interpretation of complex data must be developed.

3. Utilisation of prosodic features in information retrieval

Compared with information retrieval for text-media, semantically rich content-based information retrieval for audio/video is still an unattained goal (Bregman 1990, Smeaton 2000). With modern text mining methods, it is possible to automatically categorise and summarise very large documents, and locate and extract the relevant textual information. Image, video and audio, by contrast, are so rich in content that such generic information retrieval applications are still far away. However, some progress has already been made.

The development of automatic speech recognition can offer a partial solution to the problem of information retrieval from large and semantically rich (audio) databases: it enables the computation of topic-related keywords, thus locating those specific parts of the file where the interesting and relevant theme-related words occur. The basis of all speech retrieval applications is some kind of recognition (where phone recognition precedes word recognition). *Full speech recognition* transcribes spoken utterances into a text which can be analysed to achieve a syntactic and semantic description of the utterances, whereas *single word recognition* only locates the relevant "spots" in the spoken messages. Full speech recognition is speaker-dependent and expensive, and requires training. Word spotting applications are often sufficient and they work because of the pre-defined vocabulary (the search is for these words only).

Speech signal interpretation can be enhanced by means of prosody-based search features: general prosodic features of speech can be utilised to automatically chart the distribution of cohesive paragraphs, also known as *paratones*, in the speech data. Prosodic features such as F0 and intensity level can be used for topic and phrase boundary identification (Swerts and Ostendorf 1997). The speech can be segmented into audio paragraphs/paratones, using acoustic/prosodic information, after which automatic speech recognition can be applied to each such unit. By means of the more intelligent search features, the user of the information retrieval application could find all those lexically/syntactically relevant, and prosodically cohesive, parts of the audio which deal with a certain pre-defined topic.

To improve the current content-based information retrieval methods, it may be necessary to add prosody-based elements to the list of search features. Prosody signals not only the textual structure of the discourse – for example, the beginnings and endings of topics, and turn-taking – but also the affective state of the speakers. Speech is, in actual fact, quite rich in different kinds of *indexical* markers relating, for example, to the affective/medical status or the socio-economic/professional background of the speaker (Laver 1994). These features are, to some extent at least, speaker- and language-independent (Toivanen 2001).

Prosodic cues related to emotional content can be analysed at many different but intersecting levels by taking into account, in particular, the time-length of the speech segment. On the one hand, emotions can be detected with some accuracy in very brief speech segments/syllables of 200 ms in which the acoustic expression is, of course, entirely based on (paralinguistic) features of voice quality (Laukkanen et al. 1997). Such short units of emotional expression could be called affect bursts or semiautomatic biomechanical responses. On the other hand, phonological intonation contour types, as

well as different types of accentuation, which signal affective states at the level of the utterance or speech event, may be more cognitively mediated, and more linguistic, components of the expression of emotion.

At the *signal* level, a number of continuous variables convey emotions and affect. These variables are F0-related, intensity-related, temporal, and spectral features of the speech signal, including, for example, average F0 range, average RMS intensity, average speech/articulation rate, and the proportion of spectral energy below 1,000 Hz. These features, most of which are easily measurable from the signal, yield information about the *paralinguistic* aspects of speech. At the *symbolic* level, the distribution of tone types and focus structure in different syntactic structures can convey emotional content. Typically, these features convey *linguistic* information about how emotions are expressed in spoken language. In the research, the analysis of both of these two levels will eventually be necessary.

The kind of prosodic information briefly described above could be valuable for database indexing and retrieval of audio material. Currently available search engines cannot make use of the acoustic parameters underlying "angry" voice, for example, but more intelligent robots could be designed to utilise acoustic information. As things stand now, prosodic features are an unexplored area in research on content-based information retrieval.

4. Emotional speech databases

The computer, or the search engine, can evaluate the emotional content of spoken messages on the basis of prosodic/acoustic information only if the computer is explicitly taught the way in which human listeners perceive emotions in speech. Thus, in order to find out the quantifiable relationship between prosodic features of speech and perceived emotions, listening experiments and subjective evaluations are necessary.

Studies of emotional speech features are often based on recordings of professional actors simulating a number of (basic) emotions in a studio environment. The researchers then have a panel of listener- judges label the emotional content of each speech sample. Usually, the speech consists of a few standard phrases or sentences, the lexical content remaining the same for each simulated emotion. If the listener-judges can "hear" the intended emotions, it can be argued that the recordings are representative of the affective states under investigation (Iida et al. 1998). Because the speech material is identical, a detailed study of the acoustic/prosodic structure is justified: as the lexical content remains the same, the (spectral) differences between the speech samples, for example, should be principally caused by the intended emotional state, not by the varying sound structure.

It is difficult to find authentic emotional speech data. On the one hand, if the recording is made in a ("genuine") noisy and uncontrolled real-life situation, the stored speech signal is likely to be weak or distorted, which largely excludes a principled acoustic analysis of the data. On the other hand, ethical considerations are an important issue: is it morally acceptable, or even completely legal, to make the speaker, in a laboratory environment, genuinely angry, sad or scared for the purposes of collecting speech material?

Certain previously existing speech data can be useful for research on the acoustic correlates of emotions. The radio news broadcast of the *Hindenburg* crash is perhaps the most famous publicly available emotional speech source. Similarly, recordings made in the context of interactive radio programs can contain interesting emotion-laden speech data. The problem is, however, that the researcher cannot verify which emotion the speaker intended to produce. Furthermore, copyright restrictions often apply to such data: one cannot take for granted that the data can be accessed even for strictly limited research purposes. Indeed, most of the emotional speech databases are not available for distribution because of copyright restrictions.

5. Research utilising the MediaTeam Emotional Speech Corpus

5.1. Speech data

The *MediaTeam Emotional Speech Corpus* is currently the largest database of emotional speech for continuous spoken Finnish. Fourteen professional actors, eight men and six women, aged between 25 and 50, were recruited to simulate basic emotions. Each speaker was asked to read out a Finnish passage of 120 words in a "neutral" or "natural" tone of voice; the text was semantically as neutral as possible, dealing with the nutritional value of crowberry. Then each speaker was asked to read out the text simulating the following emotions: "happiness", "sadness", "fear", "anger", "boredom" and "disgust". The speakers were allowed to retake the reading if they were not satisfied with the first rendition. In addition, the speakers acted out two pre-written dialogues containing emotional lines of

varying length. Thus the corpus contains linguistic units with specific emotional content ranging from short exclamations to monologues of approximately 30 seconds. The speech material was digitally recorded with DAT in an unechoic studio to produce a 44.1-kHz, 16-bit CD-format recording.

5.2. Acoustic analysis

The acoustic analysis was carried out by means of analysis software, the *F0 Tool*, implemented in the MATLAB language. The software first distinguished between the voiced and voiceless parts of the speech signal (Ahmadi and Spanias 1999) and then determined the F0 contour for the voiced parts of the signal (Titze and Haixiang 1993). The F0 contours were determined by using cepstral analysis and interpolating waveform matching.

Some 40 acoustic/prosodic parameters were automatically computed from the speech signal (see Appendix for a complete list of parameters). The parameters were F0-related, intensity-related, temporal and spectral features (below, the term "segment" refers to a part of the signal of varying duration, which may be realised as silence or as voiced or voiceless speech; thus the term does not describe any linguistic/phonological unit). The most significant parameters are listed below.

The *general F0-based parameters* were the following: mean F0, median F0, maximum F0, minimum F0, F0 range, $95^{th}/5^{th}$ fractile of F0, the range between the fractiles, and F0 variance. The *parameters concerning the "dynamics" of F0* were: average F0 rise/fall during a continuous voiced segment, average steepness of F0 rise/fall, maximum F0 rise/fall during a continuous voiced segment, maximum steepness of F0 rise/fall, and jitter. The *intensity-related features* were: mean RMS intensity, median RMS intensity, maximum RMS intensity, minimum RMS intensity, intensity range, $95^{th}/5^{th}$ fractile of intensity, the range between the fractiles, intensity variance, and shimmer. The *temporal parameters* were: average duration of voiced segments, average duration of voiceless segments shorter than 500 ms, average duration of silences/pauses shorter than 400 ms, average duration of voiceless segments longer than 500 ms, average duration of silences/pauses longer than 400 ms, maximum duration of voiced segments, maximum duration of voiceless segments, maximum duration of silences/pauses, silence-to-speech ratio and voicing-to-pauses ratio. Finally, the *spectral features* measured concerned the proportion of low-frequency energy (<500 Hz and <1000 Hz).

5.3. Recognition and classification of emotional content of speech

All the speech data was tested with listener-judges: 20 listeners, female university students aged between 19 and 28, listened to each speech sample and chose the emotional label which best described the affective content of the sample. In the test, the emotions to be differentiated were: neutral, anger, happiness and sadness, which can be considered to represent the basic emotions (Izard 1977). On an average, the listeners recognised the emotions with an accuracy of 84.7%.

In the automatic recognition of emotions, k-NN (k-Nearest Neighbour) was used as a classifier. In the first experiment, the assumption was that the speaker had been authenticated earlier and the new emotional speech samples were compared with the corresponding speech models in the database. This experiment scenario was tested statistically with leave-one-out method. The automatic recognition of emotion was quite satisfactory, the average classification level being about 80 % (k-NN with k equal to 1). Further experiments are now being carried to investigate the accuracy of completely speaker-independent automatic classification of emotions (where the emotional speech samples of a person not included in the database are classified).

The results of the listening tests (>84 % for the four basic emotions) are excellent in the light of the results reported in the literature. According to Scherer et al (2001), in the "Western cultural context" listeners can recognise basic emotions (neutral, anger, fear, happiness and sadness) on the basis of acoustic features with an average accuracy of 66 %. However, if more emotions are to be recognised, the task become more difficult: an accuracy of 60 % has been reported for anger, happiness, sadness, fear, disgust, love, pride and jealousy (Scherer 1995). Polzin (1999) reports a general rate of 70 % for recognition of emotion in English (sadness, anger and neutral). The conclusion is that the speech samples used in the present investigation contained clear acoustic cues related to emotional content as the recognition rate was so high. On the other hand, it should be pointed out that the speakers were professional actors: thus powerful vocal portrayals of emotions were to be expected.

The preliminary results of the automatic recognition and classification of emotions seem very promising. Summarising research on the topic, Bosch (2000) argues that an accuracy of 60 per cent may be the best one can get in a (speaker-independent) limited happiness/joy, anger, sadness/grief discrimination test. McGilloway et al. (2000) suggest that 50 % correct classification can be achieved in automatic systems discriminating among five emotional states. Again, it must be emphasised that the

nature of our data was probably conducive to good classification results: the speech samples contained acoustic cues which the speech analysis algorithm could extract and which were realised with clearly differentiated values for different emotional states. Furthermore, it should be pointed out that the successfulness of the *speaker-independent* classification of emotion is yet to be confirmed.

5.4. Linguistic analysis

Although it is true that the acoustic parameters described in section 5.2. reveal something about the general prosodic patterns in the speech data, a more linguistically oriented analysis is needed to investigate speech melody or intonation. The acoustic parameters can be measured completely independently of the linguistic/phonological structure of the speech data: for example, the parameters describing the rate of change ("steepness") of the F0 movement do not indicate whether the most dynamic F0 contours are connected with sentence accent or not. Thus, automatic measures of mean F0-based values for the whole speech material do not provide information on the course of pitch at the utterance level, and they certainly do not provide any direct information on the linguistically relevant variation in pitch. To investigate speech melody, auditory analysis/annotation is necessary. The relation between syntax and intonation is the obvious place to start: it can be investigated whether statements, for example, contain a greater proportion of non-falling tones in certain emotional states. Our basic hypothesis is that emotional colouring might, in some cases, counteract the normal tendency for Finnish utterances to end with falling intonation. All in all, the speech material must be analysed in terms of linguistic/phonological features and units to find out how the functionally important parameters that are not directly measurable from the acoustic signal reflect different emotions. So far, the data has been analysed only in terms of automatically measurable acoustic parameters.

The distribution of sentence accents and pauses is probably very important. It can be assumed that certain emotions – anger, for example – might be accompanied by a high percentage of strongly accented words. To chart the distribution of different types of accents in the speech data, the corpus must be transcribed manually. Similarly, frequent pausing is probably to some extent an emotion-specific feature of speech: a low speech-to-silence ratio can be an indication of "passive" emotions, e.g. sadness and boredom, but the type of pause may also be important. Unlike the speech-to-silence ratio, the percentages of different types of pauses are not automatically measurable. Therefore, auditory interpretation and manual measurements are needed to investigate the locations and types of pauses. In the analysis, it is thus important to come up with some kind of (syntactic/pragmatic) taxonomy of pauses, in addition to measuring e.g. the pause/speech ratio.

To describe these features of the speech data, the corpus is currently being annotated utilising *Transcriber*, a tool for assisting the manual annotation of speech signals: the tool is especially handy as it has a graphical user interface for segmenting long recordings and transcribing them. The aim is to create interlinear transcription, where each word is annotated with prosodic and syntactic information (displayed under the word). If a piece of annotation is clicked on, the corresponding extent of audio signal is highlighted. As the recording is played back, the corresponding annotation sections are highlighted (waveform as well as text). The annotation records are stored in a spreadsheet.

The data is being annotated in accordance with the following principles. For intonation, the ToBI model is used. ToBI (*Tones and Break Indices*) is a framework for developing generally agreed-upon conventions for transcribing the intonation and prosodic structure of spoken utterances in a language variety. Originally developed for English, the system has been used in the description of the prosody of a number of languages, including Dutch, Spanish, Italian, Japanese and French – to date, the ToBI framework has not been extensively applied to Finnish. In the ToBI system, the F0 contour is associated with textual information. The system includes four tiers: the ortographic tier (ortographic form of words), the break-index tier (a 5-step scale describing the nature of breaks in speech), the tone tier (including e.g. interphrasal accents and boundary tones), and the miscellaneous tier (for extra-linguistic information, for example).

In the annotation used in this investigation, Välimaa-Blum's suggestion (1993) is followed: two complex accents, L+H* and L*+H, two boundary tones, H% and L%, and two initial boundary tones, %H and %L, are recognised. In spoken Finnish, according to Välimaa-Blum, the neutral prosodic contour is a gradually declining pattern of L+H* accents, with a low boundary tone (L%) in final position. The L*+H accent, involving late accentuation instead of the normal lexical stress on the first syllable in a word, is assumed to be typical of emotional speech acts: the entire utterance may consist of emphatic accents with delayed F0 peaks (Iivonen 1998).

In addition to the ToBI system, prosody is annotated using the "traditional" contour analysis framework. At the clause/sentence level, the intonation is transcribed as representing one of the following prosodic options suggested for Finnish by Iivonen (1998): *non-emphatic* pattern, *expressive*

pattern, and *progredient* pattern. The non-emphatic intonation is the basic pattern used in non-affective utterances, consisting of a descending F0 curve with rising-falling peaks on the accented syllables (i.e. the declining succession of L+H* accents in Välimaa-Blum's description). The expressive intonation involves a clear ascending F0 curve (at the end of a clause/sentence): in Finnish, this may indicate surprise, appeal or any clear emotional reaction. The progredient intonation is sometimes referred to as "comma intonation" (non-terminal intonation) in Finnish, involving a relatively high level of F0 before the boundary, in comparison with the low terminal intonation at the end of a declarative. Thus, while the non-emphatic pattern often involves laryngealisation (creak) with a very low F0 value in final position, the progredient intonation is typically completely periodic throughout.

Each word is annotated with respect to accentuation/prominence applying the descriptive system used by Suomi et al. (in press). It has been assumed that spoken Finnish contains at least three degrees of accentuation: words are unaccented, moderately accented or strongly accented. In the investigation carried out by Suomi et al., the tripartite division turned out to be perceptually and phonetically valid: in listening experiments only three degrees of accentuation could be reliably distinguished. Phonetically, the differences were quite clear. Mere word stress (in unaccented words) was not signalled tonally (i.e. in terms of F0), whereas moderate accent and strong accent were signalled mainly tonally. Durationally only strong accent differed from the other degrees of prominence. It seems very likely that our speech data can be annotated in terms of accentuation by means of this descriptive framework.

Finally, the breaks or pauses are annotated. The pauses are transcribed in terms of both phonetic and syntactic information: thus both manual phonetic measurement and syntactic analysis must be carried out. In the annotation, a silent pause is defined as a segment characterised by the absence of energy in the speech signal, with a minimum duration of 100 ms. Briefer pauses are not transcribed. Very brief pauses (50-100 ms in duration) typically occur between words, and may be caused by segmental features, such as plosives (Fant et al. 1990). A filled pause, with a minimum duration of 100 ms in our annotation, typically contains vocalisation of some kind. The pauses are classified along the following lines. At the first level of description, pauses can be sentence-internal (filled or unfilled) or sentence-final (filled or unfilled). At the second level of description, the (filled or unfilled) pauses may be syntactically motivated or unmotivated: syntactically motivated pauses occur at syntactic junctures (typically, between clauses and after long noun phrases). Finally, filled pauses may be "vocalisations" (e.g. *uh*, *um*) or "vocal noises" (inhalations, exhalations, laughter, sobbing, etc.). A basically linguistic approach is needed to annotate the speech corpus in terms of these features. The preliminary results indicate that the distribution of pauses signals some affective states more reliably than do most F0-related parameters: syntactically unmotivated long pauses (>500 ms) seem to be characteristic of speech reflecting sadness and, and syntactically unmotivated brief pauses (<200 ms) are connected with fearful speech.

The aim of the linguistic annotation is to complement the acoustic/prosodic analysis. We aim to find perceptual contrasts in terms of tone and accent, and see how these "quantal" properties of speech convey emotional meaning (cf. Stevens 1972), over and above the continuously variable acoustic properties. In speech, cues related to emotional content interact in layers, none of which is, of course, entirely specific to emotional expression: for example, phonological prosodic contrasts signal syntactic and informational functions, and speech rate and average intensity may be connected with the stylistic features or genre of the speech situation. The listener will hear emotional cues smeared together at different levels, which also the analytic framework should reflect.

A particularly interesting question is the time unit within which different emotional cues can be perceived in speech. It seems likely that emotional valence (i.e. whether the emotion is generally "positive" or "negative") can be accurately perceived within seconds but a more nuanced analysis would seem to require a longer time. It can also be hypothesised that phonetic features of speech convey emotional content in very brief expressions while phonological prosodic contrasts can be perceived only in longer units of speech. These are all questions which the future research will address.

6. Concluding remarks

It is already clear that the speech corpus is useful in both basic and applied research on the vocal expression of emotions in Finnish. The results are readily applicable when developing content-based information retrieval methods for (Finnish) audio databases. Furthermore, especially the linguistic analysis of emotional speech outlined above significantly increases our understanding of an important (but underexplored) aspect of spoken Finnish. This knowledge is useful, for example, when creating a new *Descriptive Grammar of Finnish* – a topical issue in Finnish studies these days.

The current speech material may represent unusually pure and intense emotional speech as the data was produced by professional actors. Thus, in the listening test, the listener-judges, though they, in the main, easily identified the basic emotions, were probably aware that the emotions represented simulated affect, differing (also prosodically) from real emotional speech. However, at the present stage of the research, this is not necessary a problem for information retrieval applications because much of the speech material stored in digital databases is, in fact, acted (radio plays, audio clips from movies, etc.). Yet it is clear that, eventually, more authentic data will be necessary. There are at least two ways to make the speech material more natural. Firstly, it is possible to use so-called *emotive texts*: the test subjects read out a text (a lengthy monologue, for example) which is emotionally biased. That is, the lexical content of the text is in tune with the intended emotion. In this way, instead of *simulating* the emotion to produce speech material, the emotion is *stimulated* or prompted, and the speech is then recorded. Secondly, genuinely authentic emotions can be induced – in the field of psychology of emotional behaviour, there are some widely used methods for evoking different emotions in the subjects as well as measuring different types of behavioural and physiological indices known to correlate with these emotions. In this type of research, the focus will eventually be geared towards the psychology of emotion, rather than the phonetic or linguistic basis of the outward manifestation of emotion.

Appendix: Prosodic parameters measured from the speech signal

MEAN	Mean F0
MEDIAN	Median F0
MAX	Maximum F0
MIN	Minimum F0
RANGE	F0 range
FRACMAX	95 % fractile of F0
FRACMIN	5 % fractile of F0
FRACRANGE	5 % - 95 % F0 range
GDPOSAV	Average F0 rise during voiced segment
GDNEGAV	Average F0 fall during voiced segment
GDRISEAV	Average steepness of F0 rise
GDFALLAV	Average steepness of F0 fall
GDRISEMAX	Maximum F0 rise during voiced segment
GDFALLMIN	Maximum F0 fall during voiced segment
GDMAX	Maximum steepness of F0 rise
GDMIN	Maximum steepness of F0 fall
NORM FREQVAR	Normalised segment frequency distribution width weighed sum
F0VAR	F0 variation
JITTER	Trend-corrected mean proportional random F0 perturbation
MEANINT	Mean RMS intensity
MEDIANINT	Median RMS intensity
MAXINT	Maximum RMS intensity
MININT	Minimum RMS intensity
INTRANGE	Intensity range
FRACMAXINT	95 % fractile of intensity
FRACINTRANGE	5 % - 95 % intensity range
NORM INTVAR	Normalised segment intensity distribution width weighed sum
INTVAR	Intensity variation
SHIMMER	Trend-corrected mean proportional random intensity perturbation
MAVLNGTH	Average duration of voiced segments
MANLNGTH	Average duration of voiceless/silent segments shorter than 500 ms
MASLNGTH	Average duration of silences shorter than 400 ms
MLNLNGTH	Average duration of voiceless/silent segments longer than 500 ms

MLSLNGTH	Average duration of silences longer than 400 ms
MAX VLNGTH	Maximum duration of voiced segments
MAX NLNGTH	Maximum duration of voiceless/silent segments
MAX SLNGTH	Maximum duration of silent segments
VRATIO	Voicing-pauses ratio
SRATIO	Silence-speech ratio
LFE500	Proportion of energy below 500 Hz
LFE1000	Proportion of energy below 1000 Hz

References

Ahmadi S, Spanias A 1999 Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Transaction on Speech and Audio Processing* 7: 333-338.

Bosch L 2000 Emotions: what is possible in the ASR framework. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast, pp 183-188.

Bregman A 1990 *Auditory Scene Analysis*. Cambridge, Massachusetts, MIT Press.

Damasio A 1994 *Descartes' Error*. New York, The Grosset Putnam.

Fant G, Kruckenberg A, Nord L 1990 Acoustic correlates of rhythmical structures in text reading. In Wiik K, Raimo I (eds), *Nordic Prosody V: Papers from a Symposium*. Turku, University of Turku Press, pp 70-86.

Iida A, Campbell N, Yasumura M 1998 Design and Evaluation of Synthesised Speech with Emotion. *Journal of Information Processing Society of Japan* 40: 479-486.

Iivonen A 1998 Intonation in Finnish. In Hirst D, Di Cristo A (eds), *Intonation Systems: A Survey of Twenty Languages*. Cambridge, Cambridge University Press, pp 311-327.

Izard C 1977 *Human Emotions*. New York, Plenum Press.

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

Laver J 1994 *Principles of Phonetics*. Cambridge, Cambridge University Press.

McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S 2000. Approaching automatic recognition of emotion from voice: a rough benchmark. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast, pp 200-205.

Polzin T S 1999 *Detecting Verbal and Non-verbal Cues in the Communication of Emotions*. Unpublished PhD thesis, Carnegie Mellon University.

Scherer K R 1995 Expression of emotion in voice and music. *Journal of Voice* 9: 235-248.

Scherer K R, Banse R, Wallbott H G 2001. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology* 32: 76-92.

Smeaton A 2000 TREC-6 Personal Highlights. *Information Processing and Management* 36: 87-94.

Stevens K N (1972) The quantal nature of speech: evidence from articulatory-acoustic data. In David E, Denes P (eds), *Human Communication: A Unified View*. New York, McGraw Hill, pp 51-56.

Suomi K, Toivanen J, Ylitalo R (in press) Durational and tonal correlates of accent in Finnish. *Journal of Phonetics*.

Swerts M, Ostendorf M 1997 Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication* 22: 25-41.

Titze I R, Haixiang L 1993 Comparison of F0 extraction methods for high-precision voice perturbation measurements. *Journal of Speech and Hearing Research* 36: 1120-1133.

Toivanen J 2001 *Perspectives on Intonation: English, Finnish and English Spoken by Finns*. Frankfurt am Main, Peter Lang.

Välimaa-Blum R 1993 A pitch accent analysis of intonation in Finnish. *Ural-Altaische Jahrbucher* 12: 82-89.