EMOTIONS:
WHAT IS POSSIBLE IN THE ASR FRAMEWORK

Louis ten Bosch,
louis.tenbosch@lhs.be

Lernout & Hauspie Speech Products N.V.

ABSTRACT

This paper discusses the possibilities to extract features from the speech signal that can be used for the detection of emotional state of the speaker, using the ASR framework.

After the introduction, a short overview of the ASR framework is presented. Next, we discuss the relation between recognition of emotion and ASR, and the different approaches found in the literature to tackle the correspondence between emotions and acoustic features. The conclusion is that emotion itself will be very difficult to predict with high accuracy, but in ASR general prosodic information is potentially powerful to improve the (word) accuracy for tasks on a limited domain.

1. INTRODUCTION

'Emotion in speech' is a topic that receives much attention during the last few years, both in the context of speech synthesis as well as in automatic speech recognition. The advantage of 'emotionally rich' speech synthesis is evident. The approach to simulate the effect of emotion in synthetic speech is usually based on acoustic analyses of databases of (human) 'emotional' speech. These databases contain utterances spoken by actors in certain emotional ways. From these studies, it appears that a number of basic emotions such as anger, sadness and happiness can quite well be described in terms of changes in pitch, duration and energy [e.g. 10, 20, 22], of which pitch is the most important [26]. Modification of these parameters also shows good results in speech synthesis for different languages (see e.g. [12, 19]).

Also for ASR, the recognition of emotion in speech can be useful. However, the automatic detection of emotion from the speech signal is not straightforward. For example, it appears from recent studies that the triplet happiness, sadness/neutral and anger can be detected only with an accuracy of 60-80 percent [20, 26]. When more emotional 'modes' are to be recognized (some studies distinguish 8 or more different emotions), the detection score decreases substantially. A number of studies narrow down 'emotion' to 'stressed', in the sense of stressful (e.g. [5]). The task is to binary classify utterances as 'stressed' or 'not stressed'. With optimal choice of the features used for classification, stress detection results approach or exceed 90 percent. Other studies narrow down the rather broad concept of emotion to a number of more pragmatic classes: for example approval, attention, and prohibition in parent-child interactions [7]. This study shows a speaker-independent recognition score of about 55 percent, and, interestingly, a large speaker dependency of the accuracy ranging from 60 to 90 percent (after a speaker dependent training of the classifier).

Most of these studies concerning the detection of emotional 'state' of the utterance do not bother about the linguistic (text) content of the utterance. They mainly deal with 'stand alone' recognition modules for emotion. In the following section, we will discuss in more detail how emotion can be used in the ASR framework.

2. THE ASR FRAMEWORK

In this section we briefly discuss the mostly used paradigm for automatic speech recognition. Most of today's automatic speech recognition (ASR) systems treat the speech signal as an example of a stochastic pattern and use statistical pattern recognition techniques to produce a word sequence hypothesis. Speech recognition is mostly defined as solving a maximum a posteriori (MAP) problem, in which, for a incoming sequence of acoustic vectors A, a sequence of words W must be found such that

$$P(W \mid A)$$

is optimized, usually under additional constraints imposed by a grammar. Under the general assumption that the Bayesian rule applies, we obtain $P(W \mid A) = P(A \mid W) * P(W) / P(A)$, and so (P(A) being fixed):

$$\text{argmax}_W P(W \mid A) = \text{argmax}_W \{ P(A \mid W) * P(W) \}$$

We see that we basically need the evaluation of two factors, the first being $P(A \mid W)$, which is the probability of observing a sequence of acoustic vectors given the word sequence, and the second factor P(W), denoting the probability of the word sequence itself. The first probability relates to the acoustic model, while the second probability is referred to as language model. For a commercial dictation system, the acoustic model (AM) is usually trained using an acoustic training database of 50 to 150 hours of speech, while the language model (LM) may require a text corpus containing 100-1000 million words. The algorithm that actually performs the optimization of $P(W \mid A)$ is based on a pattern recognition approach, and is often implemented by using dynamic programming (DP) techniques.

The sequence of acoustic vectors A is the result of a properly chosen feature extraction algorithm (FE). Two properties of the FE are relevant for the discussion here. Firstly, the FE produces a sequence of feature vectors or frames, typically 100 a second, each vector representing the spectral characteristics of the sound at that moment. Due to the windowing, the vector represents the

'average' of the speech spectrum over about 20-30 ms. Therefore, the features contain information that is very local in time. Secondly, for regular speech recognition tasks, the FE is to be designed to normalize for all effects that have nothing to do with the linguistic content. Such effects include background noise, undesired channel effects, line echoes, microphone characteristics, but also speaker-dependent characteristics such as the vocal tract length. Speech features that are commonly left out or at least neglected by the FE are the pitch and duration. (The pitch is not left out in a number of ASR systems designed to recognize tonal languages such as Chinese).

The feature extraction can be implemented in many ways, but a very common, if not quasi-standard way, is to use mel-based cepstral coefficients. These are based on an (fast) fourier transform, followed by a non-linear warp of the frequency axis, the logarithm of the power spectrum, and the evaluation of the first N coefficients of this log warped power spectrum in terms of cosine basis functions. (In recent years, other promising approaches get the attention that they deserve, such as auditory features, articulatory features, and discriminative features.) The general focus of the FE is to produce short-time spectral features that are normalized for a number of factors that are considered irrelevant for the 'text' content of an utterance. The way an utterance is produced, with high or low pitch, fast or slow, angry or sad, these are all irrelevant factors as seen from the default ASR point of view.

Another well-known aspect of the ASR 'engine', which is of ultimate importance for the modeling power of a recognizer, is the use of so-called Hidden Markov Models (HMMs). The commonly used algorithms all use HMMs to model the acoustic properties of the recognition tokens. These recognition tokens may be entire words, or (mono)phones, diphones, triphones, multiphones, syllables, multi-words, or combinations of them, etc. No matter how an ASR algorithm is designed, sooner or later there has to be some construction to connect the incoming speech vectors (from the FE), the information from the lexicons and the language model. The final output of the ASR algorithm, often a list of N-best hypotheses, is then the result of competition between acoustic and language scores, from the acoustic model and the language model, respectively. The issue here is, that in the standard manner of training an HMM model, all HMM states with be aligned with a small number of acoustic frames, thereby modeling a short acoustic event. So, although the HMM models themselves can be used to model speech units of various lengths, the HMM states 'in the classical' ASR still correspond to small time scale events, on segmental or maybe syllable level. Emotion, however, is a feature that manifests itself beyond syllable level.

The last aspect that we want to emphasize is the Markov assumption. This assumption is applied in the evaluation of P(W) as well as inherently used in the evaluation of P(A ¦ W) in evaluating the HMMs. The fact that commercially available recognition systems perform reasonably well (with a score of 90-95 percent for a dictation task using a lexicon of 60000 words, after speaker adaptation of the acoustic models) shows that the Markov assumption is not that bad. But from a theoretical point of view, the assumption does not hold, and it is to be considered as a drastic simplification of reality. Whether features such as emotion can be modeled using the Markov assumption is not straightforward.

Evidently, the 'linguistic content' of an utterance goes beyond the concatenation of units on segmental level. The use of pitch to mark prominence, or to disambiguate meaning, or to put emphasis on parts of speech in focus, or the use of volume to attract attention are examples of this. According to many studies, pitch is the most relevant acoustic parameter for the detection of emotion, followed by duration and energy. When we want to integrate supra-segmental information, say pitch, into the ASR/HMM paradigm, there are principally two methods, one related to the 'front end' of the recognizer, the second one to the 'back end'.

1. In the first method, the FE is enriched with a pitch detection algorithm. The pitch feature (or a related feature describing the harmonic structure of the spectrum) is used to create a separate 'acoustic models' (using e.g. a pitch/delta-pitch codebook). The acoustic model used in tests is a combination of the gross-spectral model and the pitch model. In this way, one can improve the performance of a recognition system for Chinese with about 10-30 percent reduction in word error rate. The 'tones' are to be transcribed in the lexicon on the syllable level.

2. The second method is to evaluate the pitch in the FE, but to use the pitch information only at the stage of rescoring the N-best list that is produced by the regular ASR recognizer. In this way, one may use pitch, word stress or other supra-segmental or lexical information to improve the recognition score [see e.g. 21]. In contrast with method 1, this method can also be used when pitch is to mark information that is on the supra-word level, e.g. focus. Therefore this or a similar method is particularly useful in improvement of recognition results in dialogues. We come back to this method in the next section.

As an illustration of focus in a dialogue system, the simple negation

No, I'll take the train to London at 5 p.m.

has a number of readings depending on the pitch contour. The use of prosodic information is also of ultimate importance for studying the dialogues of players talking within a limited discourse domain. The success so far has been limited, but recently more progress has been claimed in the relation between ASR performance and prosodic properties of utterances [compare e.g. 2, 3 and references herein].

3. EMOTION AND ASR

Human emotions include love, sadness, fear, anger, and joy/happiness as basic ones, and some people add hate, surprise, and disgust, and distinguish 'hot' and 'cold' anger. Some studies distinguish 'emotion' from 'mood': an emotion is always referring to an object: one grieves over something, one loves somebody, etc. In this more precise sense, we here deal with the reflection of mood, rather than of emotion, in the speech signal. (We will stick to the word emotion, though.)

3.1 Affective computing

Quite some research effort is now being put into a field what is called 'affective computing' [9]. The goal is to design ASR and TTS related algorithms that understand and respond to human emotions. The commonly applied approach is to start with a database with emotional that is annotated with emotional tags by a panel of listeners. In most cases, such utterances are spoken by actors mimicking specific emotional states (see e.g. [6, 7, 17, 19]). The next step is to perform an acoustic analysis on these data, and to correlate statistics of certain acoustic features (pitch, pitch range, etc.) with the emotion tags. This step may involve techniques also used in ASR (such as Gaussian modeling), but also other classification techniques are used (VQ, ANN) [26]. In the third step, the resulting parameter estimates are verified and adapted by using a speech synthesis tool, and by a final check of the synthesis in a human classification test.

In the context of ASR, an appropriate way to deal with emotion is to regard an utterance on a number of levels:

- text (segmental) level. This level is accessed by the classical ASR - the voice activated typewriter.
- prosodic (supra-segmental) level: pitch, volume, pausing, phrasing, speaking rate: this level is addressed by recognition systems aiming at the transcription/detection of word accents, intonation patterns, pausing.
- 'emotion' level: neutral, sadness, happiness, anger, etc.
- 'functional' level: directive, question, approval, attention, prohibition, etc.

There is an interaction between ASR performance and prosodic properties of the utterance. First of all, ASR performance is known to vary depending on the level of formality and speaking style [14, 15]. For most systems, in order to obtain optimal ASR result, the ASR test conditions should be 'the same' as the ASR training conditions, and so variations in speaking style and speaking rate have a negative impact for the ASR performance. To speak slower than normal is usually less worse than speaking faster than normal. It is a well-known effect that customers of a voice operated information system or IVR system tend to hyper-articulate when they can't get through the dialogue, which is usually a bad strategy to get better recognized (see e.g. [1]). But prosody can be also used in a positive way. For example, a number of studies show that prosody itself is capable of re-ranking the ASR hypotheses such as to separate the correctly recognized utterances from incorrectly recognized ones [11, 4, 2, 3]. In fact, it is claimed in [3] that some prosodic features can more accurately predict when an ASR hypothesis contains a word error than acoustic confidence scores do. That means that some prosodic features provide useful information to explain ASR recognition failure. It is not yet clear [3] whether these features directly hamper the ASR search (and therefore trivially correlate with word recognition errors) or whether they are indirectly associated with properties in the speech signal that deteriorate ASR performance.

ASR errors can often, but not always, be associated with prosodic effects in the speech signal, mainly with speaking rate, and phrasing. Although ASR systems are designed *not* to be sensitive to pitch and loudness variations, these variations can still percolate though the feature extraction and affect the acoustic modeling and the test. In other words, the output of the ASR recognition may depend on prosody of the speech input, although the ASR only worries about the words – in that view, the impact of prosody is just a negative side effect. This dependence is undesired or at least not aimed at for the classical ASR, but may at the same time be useful and advantageous for ASR systems serving e.g. a dialogue purpose.

In general, the ability to generate speech with a particular emotional value does not at all guarantee the ability to correctly recognize that emotion from the speech signal. Broadly stated, synthesis just needs one good exemplar, which the ASR must be robust against the whims of fashion of the speaker.

3.2 Findings in literature

Research of emotions in speech primarily deals with the search for acoustic features of speech that distinguishes a number of emotional states. The topic receives increasing attention during the recent three, four years. In this section we aim at an overview of approaches and results.

In [6] the emotions anger, fear, sadness, anxiety, and happiness were studied in terms of their prosodic and articulatory correlates. It was found that especially speech rate, segment duration and accuracy of articulation are useful parameters to determine the emotional state of the speaker. For example, sadness was clearly shown to correlate with slow speech, while fear correlates with a higher speaking rate than average. Anxious utterances show segments that are shorter than average, with exception of voiceless plosives which are often aspirated. Also in [8], relations were shown between the emotional state and the duration of vowels and consonants. But pitch is the speech feature that is most useful to distinguish emotional state [8], or anyway to convey supra-textual information. In [7], a study was conducted to automatically classify an utterance (spoken by a parent to a young infant) into three types: approval, attention, and prohibition. Compared to a system that is to detect emotional states, this looks like an easy task, but it appears far from trivial to obtain a good performance. Based on pitch slope, mean pitch and mean delta pitch, measured globally on the entire utterance, the results were close to 55 percent correct on average. To define percentage correct, the automatic

classification has been compared with some human consensus classification. One of the key observations in this study is that emotional 'production' 'varies wildly' among individuals. Classifiers that have been based on speaker dependent features showed correctness scores ranging from 60 percent up to 92 percent (based on 30 to 50 utterances per parent-infant pair). Apart from the relation between emotion and pitch, pitch range, tilt, pronunciation accuracy, also a relation between emotion and vocal quality have been claimed [23].

Synthesis. Many of the emotion studies in fact deal with the art of speech synthesis [see e.g. 12, 18, 20, 24, 25]. To simulate emotional states in synthesis, one evidently needs direct specific control especially over pitch, segment duration and phrasing parameters to create the desired emotional effect. Speech synthesis modules have been as a tool to study the impact of supra-segmental features on the perception of emotion. For ASR, however, segment duration is not an easily accessible parameter of the speech signal. Moreover, the speaking rate correlates with many more speech and speaker characteristics, e.g. with articulatory sloppiness and non-nativeness of the speaker.

Some studies attempt to synthesize emotional speech with a speech synthesizer using a parameter space covering not only f0, duration and amplitude, but also voice quality parameters, spectral energy distribution, harmonics-to-noise ratio, and articulatory precision [e.g. 18]. They focus at the four emotions anger, sadness, fear and disgust. They conclude that sadness is the most 'distinctive' emotion, compared to stimuli of each of the other emotions. In [20] one aims at recognition of seven emotions: neutral, cold anger, hot anger, happiness, sadness, interest and 'elation'. The acoustic parameters used were fundamental frequency, energy, standard deviation of energy, jitter, and shimmer; all parameters measured globally across utterances, and appropriately averaged. In this study, anger and sadness could quite clearly be distinguished from each other, but other emotions show quite a large confusability. (The database contained two speakers only – which is too small to draw firm conclusions.)

In a number of cases, synthesis model parameters are also based on rules derived from a database with speech with 'emotional prosody' (for e.g. Spanish see [12]).. Using this collected data, a rule-based simulation of three primary emotions was implemented in the TTS system. It was attempted to simulate the three emotions happiness, sadness, and anger using manipulation of pitch (range, level, slope), and a number of additional parameters (spectral tilt, and noise that is added to the voice source). The resulting success rate of about 60-70 percent. The same technique was applied for Japanese [24], in an attempt to improve the expression of the three emotions joy, anger, and sadness by using CHATR, the concatenative speech synthesis system being developed at ATR. A perceptual experiment was conducted using stimuli synthesized on the basis of each emotion corpus. The results proved to be significantly identifiable by a panel. From these prototypical databases, they study the acoustic features relevant for specifying a particular emotion. F0 and duration showed significant differences among emotion types. They showed that mean fundamental frequency was lowest for sadness and highest for happiness/joy. Duration per phone for sadness was longest and for anger was the shortest. They also looked at pauses, and the only significant finding was that pauses were longer in the 'sad' corpus than they were in the other corpora.

Influence of culture. It may be difficult to identify the emotion of a speaker from a different culture [16, 19]. In [16], it was also found that listeners will predominantly use the visual mode to identify emotion if they have the chance to do so. Cultural similarities and differences between 7 Japanese and 5 North American subjects have been compared in the recognition of emotion. Japanese and American actors made vocal and facial expression (short utterances) to transmit six basic emotions: happiness, surprise, anger, disgust, fear, and sadness. There were three presentation conditions: auditory, visual, and audio-visual. It was shown that subjects using the auditory mode can more easily recognize the vocal expression of a speaker who belongs to their own culture (the subjects were not bilingual). Both Japanese and American subjects identify the audio-visually incongruent stimuli more often by the visual mode rather than by an auditory mode.

Language dependencies. Emotional patterns may be language dependent [19]. This study examines how prosody contributes to the percept of emotions in Japanese and French synthesized speech. They find the major features determining the emotion to be pitch, speaking rate, duration and the energy of syllables. They found prosodic parameters for five emotions: anger, surprise, sorrow, hate, and joy. Responses to the synthesized speech showed that the parameters of anger, sorrow and hate are confirmed over 85 percent. Their experimental results suggest that surprise and joy may depend more on semantics, rather than prosody.

Linear – nonlinear features. Another approach is taken in [5]. Rather than the effect of emotion in general, they study the effect of a stressful situation on the speech characteristics. Stressful or highly emotional modes usually deteriorate the performance of a speech recognition system. To address this, they investigate a number of linear and nonlinear features and processing methods for the classification of what the authors call 'stressed' speech. The linear features include properties of pitch, duration, energy, glottal source parameters. The nonlinear part of the processing is based on the 'Teager Energy Operator', incorporation of frequency domain critical band filters and properties of the resulting TEO auto-correlation envelope. The TEO in discrete form reads

$$TEO(x[n]) = x[n]*x[n] - x[n+1]*x[n-1]$$

which acts like a non-linear 'energy'. The classification algorithm is based on the Bayesian hypothesis testing and hidden Markov modeling. For each stress condition, a Gaussian pdf has been modeled to match the training vectors – these training vectors were sequences of measurements of the individual features over time. The tests focuses on utterances under adverse conditions such as 'loud', 'angry', and the Lombard effect from the database SUSAS ('Speech under Simulated and Actual Stress'). This database had been exploited earlier by one of the co-authors. Results using ROC curves and

EER based detection clearly indicate that pitch is the best of the five 'linear' features for stress classification (result about 88 percent); the nonlinear TEO-based feature, however, outperforms pitch by about 5 percent. The authors observe that stressed speech seems to be affected differently across frequency bands. (In phonetic studies, similar effects are observed. It is well known that there is a relation between spectral tilt of a vowel sound and the presence of word stress on the corresponding syllable. This relation is based on the correlation between word stress, vocal energy and mouth aperture. Unfortunately, the quantification of this effect is vowel dependent to a large extent.)

This study shows the effectiveness of particular 'linear' and 'non-linear' features for detection of 'stressed speech'. For ASR this suggests that the focus should be always on formulating robust features, which are less dependent on the speaking conditions, rather than on the application of compensation or adaptation techniques.

The speech material in [13] consisted of two semantically neutral utterances spoken by two actors (one male, one female) mimicking a neutral tone and three moods: anger, happiness and sadness. The duration, fundamental frequency (F0) and the sound intensity (RMS) were used as features. Also this method showed that the fundamental frequency parameter was the most distinctive, showing differences between anger and happiness according to the shape of the contour, and between "cold" anger and "hot" anger on F0 mean. The study confirms findings showing hot anger and happiness having an large F0 range and high mean in contrast to the emotion of sadness, and the neutral voice.

Short term-long term. As we could expect, long term features seem to outperform short-term features [26]. It was attempted to recognize the emotional status of individual speakers by using speech features extracted from short-time as well as long-time analysis frames. The classification task was to distinguish 6 emotions: neutral, happiness, anger, fear, surprise, and sadness. A principal component analysis was used to analyze the importance of individual features in representing emotional categories, and to reduce the dimensionality (the number of features used in the recognition system is reduced from 22 to 12, per utterance). Three classification methods (vector quantization, artificial neural networks and Gaussian mixture density model) were used; and classifications were carried out using short-term features only, long-term features only and both short-term and long-term features. The Gaussian mixture density method with both short-term and long-term features showed the best recognition performance (62%, based on 5 speakers, 15 sentences/speaker in training, 5 in test, so also in this study the test is quite small). The analyses show that of the six emotions, there are three groups that stand out with respect to distinctiveness: neutral-sadness, anger-fear, and happiness-surprise. Within these groups, the separation is much more difficult. In another study, [17] discusses a method in which an 'emotional index' is evaluated over time, thereby avoiding the choice between short and long term features. A set of basic emotions is defined, and for each such emotion a reference point is computed. At each instant the distance of the measured parameter set from the reference points is calculated and used to compute a 'membership index' for each emotion, the 'emotional index'. In this preliminary study, the authors report success rates of about 50 percent for 5 emotions (acoustic measurements based on 24 speakers).

4. DISCUSSION --CONCLUSION

In this section, we summarize a number of issues that have been discussed or addressed on many of the cited studies, and which are all related to the possible use of emotion-related features in ASR.

1. Acoustically, emotions overlap and appear in various degrees. Of all the basic emotions the triplet happiness, anger, and sadness can be most clearly distinguished in terms of the phonetic features. In general, the most useful phonetic feature is shown to be pitch; one study [5] defines a non-linear energy-related feature outperforming pitch in a cut down stress detection task.

2. In general, the recognition of emotion is not straightforward. A score of 60 percent is about the best one can get in a limited happiness/joy, anger, sadness/grief discrimination task.

3. The acoustic realization of specific emotions seems to be speaker dependent to a large extent.

4. There is evidence that the acoustic realizations of emotions seem to be language dependent.

5. If more emotions are to be classified, one study supports a quite clear distinction between the three 'groups' neutral-sadness, anger-fear, and happiness-surprise. Within these groups, the separation is much more difficult.

6. Gaussian modelling is among the best methods to distinguish emotional classes in a space spanned by the following phonetic parameters: pitch, pitch range, average pitch, all measured across the entire utterance after endpointing (i.e. pause/speech boundary detection).

7. Almost all studies support that pitch mean is lowest for sad speech and highest for joy/happiness.

8. The validation of an automatic emotion recognition system is based on subjective judgements from a panel. A number of studies show that it is difficult to define an objective scale for subjective phenomena. That is one reason why an analysis may perform well on an individual basis for each speaker, and worse for all speakers simultaneously. Such effects seem to have played a role in many studies cited here. Automatic classifications can only be as good as the reference data. In the best case, the reference databases have been annotated by a form of consensus labeling.

9. Findings reported in the literature support the statement that prosodic information shows potential power to improve the ASR (word) accuracy for tasks on a limited domain. The same may hold for emotional information.

10. In most studies, the training and test sets are quite small – several orders smaller than the acoustic databases used in regular ASR training and test. Given the possibility of speaker dependency and the dependency of semantics, conclusions on the possibility of automatic detection of emotional tags cannot be really firmly justified in a general case.

5. REFERENCES

1. Soltau, H. and Waibel (1998), A. On the influence of hyperarticulated speech on recognition performance. Proceedings ICSLP 1998, pp. 229-232.

2. Hirschberg, J, Litman, D. and Swerts, M. (2000). Prosodic cues to recognition errors. Paper presented at ASRU 99, Keystone, USA, dec. 1999.

3. Litman, D., Hirschberg, J. and Swerts, M. (2000). Predicting automatic speech recognition performance using prosodic cues. Paper presented at NAACL in Seattle (May 2000).

4. Hirose K. (1997). Disambiguating recognition results by prosodic features. In: Computing Prosody: Computational models for Processing Spontaneous Speech, p. 327.

5. Zhou G., Hansen J.H.L., Kaiser J.F. (1998). Linear and nonlinear speech feature analysis for stress classification. Proceedings ICSLP 1998, pp. 883-886.

6. Kienast, M., Paeschke, A., Sendlmeier, W. (1999). Articulatory reduction in emotional speech. Eurospeech 1999, p. 117-120.

7. Slaney, M., and McRoberts, G. (1998) Baby ears: a recognition system for affective vocalizations. Proceedings ICSLP 1998 (on cdrom).

8. Murray, I.R. and Arnott, J.L. (1993). Towards a simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. JASA 93 (2), p. 1097-1108.

9. Picard, R.W. (1997). Affective computing. MIT Press, Cambridge MA, 1997.

10. Mozziconacci, S, Hermes, D. (1998). Study of intonation patterns in speech expressing emotion or attidtude: production and perception. IPO annual progress report, IPO, Eindhoven, 1998.

11. Veilleux N. (1994). Computational models of the prosody/syntax mapping for spoken language systems. PhD thesis, Boston University.

12. Montero J.M., Juana M. Gutierrez-Arriola, Palazuelos S., Enriquez E., Aguilera S., Pardo J.M. (1998). Emotional Speech Synthesis: From Speech Database to TTS. Proceedings ICSLP 1998, pp. 923-926.

13. Pereira, C., Watson, C. (1998). Some Acoustic Characteristics Of Emotion. Proceedings ICSLP 1998, pp. 927-930.

14. Weintrab M. et al. (1996) Effect of speaking style on LVCSR performance. Proceedings ICSLP, 1996.

15. Oviatt, S.L. (1998). The CHAM model of hyperarticulate adaptation during human-computer error resolution. Proceedings ICSLP, 1998, pp. 3211-2314.

16. Shigeno, S. (1998). Cultural Similarities and Differences in the Recognition of Audio-Visual Speech Stimuli. Proceedings ICSLP 1998 (cdrom).

17. Amir, N., Ron S. (1998) Towards an Automatic Classification of Emotions in Speech. Proceedings ICSLP 1998, pp. 555-558.

18. Rank E., Pirker H. (1998). Generating Emotional Speech with a Concatenative Synthesizer. Proceedings ICSLP 1998, pp. 671-674.

19. Koike K., Suzuki H., Saito H. (1998). Prosodic Parameters in Emotional Speech. Proceedings ICSLP 1998, pp. 679-682.

20. Whiteside S. P. (1998). Simulated Emotions: an Acoustic Study of Voice and Perturbation Measures. Proceedings ICSLP 1998, pp. 699-703.

21. Streefkerk, B.M., Pols, L.C.W, Ten Bosch, L.F.M. (1998). Automatic Detection of Prominence (as Defined by Listeners' Judgements) in Read Aloud Dutch Sentences. Proceedings ICSLP 1998, pp. 683-686.

22. Cowie, R. Douglas-Cowie, E. (1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech. Proceedings ICSLP 1996.

23. Zetterholm E. (1998) Prosody And Voice Quality In The Expression Of Emotions. Proceedings ICSLP 1998 (on cdrom).

24. Iida, A., Campbell, N., Iga S., Higuchi F., Yasumura M. (1998). Acoustic Nature and Perceptual Testing of Corpora of Emotional Speech. Proceedings ICSLP 1998, pp. 1559-1562.

25. Mizuno O., Nakajima S. (1998). A New Synthetic Speech/Sound Control Language. Proceedings ICSLP 1998, pp. 2007-2010.

26. Li Y., Zhao Y. (1998). Recognizing Emotions in Speech Using Short-term and Long-term Features. Proceedings ICSLP 1998, pp. 2255-2558.