Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 40 (2003) 213–225

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Emotions, speech and the ASR framework

Louis ten Bosch

A2RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Abstract

Automatic recognition and understanding of speech are crucial steps towards natural human–machine interaction. Apart from the recognition of the word sequence, the recognition of properties such as prosody, emotion tags or stress tags may be of particular importance in this communication process. This paper discusses the possibilities to recognize emotion from the speech signal, primarily from the viewpoint of automatic speech recognition (ASR). The general focus is on the extraction of acoustic features from the speech signal that can be used for the detection of the emotional state or stress state of the speaker.

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 that deal with the correspondence between emotions and acoustic features. The conclusion is that automatic emotional tagging of the speech signal is difficult to perform with high accuracy, but *prosodic* information is nevertheless potentially useful to improve the dialogue handling in ASR tasks on a limited domain.

Keywords: Emotion; Prosody; Automatic speech recognition

1. Introduction

'Emotion in speech' is a topic that has received much attention during the last few years, in the context of speech synthesis as well as in automatic speech recognition (ASR). Speech is the most convenient means of communication between people. Although we are still far from having a machine able to communicate with a human in a natural way, scientific and technical improvements show a direction towards a more natural man–machine speech interface and natural language processing (NLP). The use of emotion in speech synthesis, and the recognition of emotion in speech recognition can substantially contribute to the naturalness of man–machine communication.

For text-to-speech (TTS) systems, the advantage of 'emotionally rich' speech synthesis is evident (also see Murray and Arnott, 1993). The pragmatic value of TTS systems is mainly determined by two factors: the intelligibility of the speech that is produced, and the naturalness of the speech output. In the last decades, the improvement of segment intelligibility, and the smooth concatenation of the synthetis units have been a driving force in the design of TTS systems; as a result, the TTS word intelligibility has substantially improved during recent years. However, less success has been achieved in making the synthetic speech more natural. Even in modern TTS systems, there is quite a long way to go, in order to improve the

E-mail address: l.tenbosch@let.kun.nl (L. ten Bosch).

0167-6393/02/$ - see front matter
PII: S0167-6393(02)00083-3

prosody and naturalness of the synthetic signal (Juang and Furui, 2000; e.g. Chu and Peng, 2001).

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 or students who have been asked to read the same utterance simulating different emotions (see e.g. Zhao et al., 2000), or on simulated dialogues (Hirose et al., 2000), or using a Wizard-of-Oz setting (Huber et al., 2000). A large number of studies investigate the relation between acoustic features of utterances and the emotion tags given to these utterances by humans in a perceptual labeling task. It appears that a number of 'basic' emotions such as anger, sadness and happiness can quite well be described in terms of changes in prosodic factors: pitch, [1] duration and energy (e.g. Scherer, 1981; Van Bezooijen, 1984; Frick, 1985; Mozziconacci and Hermes, 1998; Whiteside, 1998; Cowie and Douglas-Cowie, 1996; Mozziconacci, 2000), of which pitch is in many cases the most important (e.g. Li and Zhao, 1998; Yang, 2000). Modification of the relevant TTS parameters also shows good emotion simulation results for different languages (see e.g. Montero et al., 1998; Koike et al., 1998).

Also for ASR, the recognition of emotion in speech can be useful, e.g. for proper handling of a man–machine dialogue. However, the automatic detection of the emotion state from the speech signal is not straightforward. For example, recent studies show that the triplet happiness, sadness/neutral and anger can be distinguished only with an accuracy of 60–80% (Whiteside, 1998; Li and Zhao, 1998). When more emotion tags are to be recognized (some studies distinguish eight or more different emotions), the detection results decrease substantially; depending on the task, the performances range from 25% to about 50%. These results are largely dependent on the size of the acoustic datasets, and on the way the emotion classifier is trained. Kang et al. (2000) obtain a high performance (higher than 80%) using 6 emotions, but this study is based on a very limited word set (8 speakers, 5 different words). Their study furthermore shows substantial differences in performance between three different classification algorithms (Gaussian modeling and maximum likelihood, nearest neighbour, and hidden Markov modeling).

A number of studies narrow down 'emotion' to 'stressed', in the sense of stressful (e.g. Zhou et al., 1998). In that case, the task is not to recognize the emotion itself, but to binary classify utterances as stressed or 'not stressed'. With an optimal choice of the features used for classification, such a stress detection may result in a recognition score of about 90% under non-adverse recording conditions (the test set consisted of 30 words). (Another study relating emotion and stress is Fernandez and Picard, 2000.) Other studies limit the rather broad concept of emotion to a number of more pragmatic classes: for example approval, attention, and prohibition in parent–child interactions (Slaney and McRoberts, 1998). This study shows a speaker-independent recognition score of about 55%, and, interestingly, a large speaker dependency of the accuracy ranging from 60% to 90% (after a speaker-dependent training of the classifier). Huber et al. (2000) simplify the classification of emotion to the binary question whether one particular emotion (anger) is present or absent. They report the highest classification (so, anger versus neutral) result of 86% for acted speech, but a more realistic test using non-acted speech (in a Wizard-of-Oz setting) yielded a classification rate of 66% correct, which shows the sensitivity of performance results on the paradigm of the test (for more details see Huber et al., 2000).

A very pragmatic issue is the *usefulness* of emotions to ease the human–machine communication. For example, in dialogue systems it may be very fruitful to be able to detect 'moods' such as frustration, irritation or impatience, such that a more appropriate dialogue handling can be chosen or the call be redirected to a human attendant. From this perspective, the emotions happiness, fear, anger themselves are much less relevant than the 'ease of communication' qualifiers such as irritation and impatience. Klein (1999) shows that a strategy focusing on ease of communication might

[1] Throughout this paper, we will use the terms pitch and fundamental frequency ($F0$) interchangeably. A number of studies reserve the term pitch for the percept related to the physical parameter fundamental frequency.

work in real life. At the MIT Media Lab, he designed a human–computer interaction agent that was built to support users in their ability to recover from negative emotion states, particularly frustration. The agent used social-affective feedback strategies, and its effectiveness was evaluated against two control conditions in a 72-subject study. Behavioral results showed the emotional agent was significantly more effective than a 'neutral' agent in helping relieve frustration levels.

The recognition of emotion is partly based on 'paralinguistic decoding'. Although the emotion may manifest itself on the semantic level, the emotion content is to an important extent carried by prosodic features. So, while ASR focuses on the correct recognition (in terms of a sequence of words) for a given acoustic input, the emotion in an utterance is mainly encoded in prosody and semantics, which are areas that are not directly focused on in the mainstream ASR approaches. Broadly speaking, the process of speech perception includes: the detection of acoustic–phonetic cues to form/activate words, a grammatical analysis to form well-formed sentences, semantic determination and disambiguation, and, on top of this, the pragmatic use of prosodic cues. ASR mainly deals with the first two stages, while emotion is mainly encoded in the second, third and forth stage, where techniques from natural language processing apply. For classical ASR systems, the most important quality measure is still the word error rate (or recognition token error rate), and the emotion in the utterance does not play a central role. For natural language understanding (NLU), or dialogue systems, a proper response needs understanding of the context of the utterance, and it is in this setting that prosody and emotion tags may play a more important role.

In the following section, we will discuss in more detail how emotion can be used in the ASR framework, and how a separate classification of emotion might be fruitfully used in this framework.

2. The ASR framework

In this section we briefly discuss the mostly used paradigm for ASR. By ASR we mean the set of algorithms that hypothesize a word sequence given an incoming acoustic speech signal. Most of today's ASR systems treat the speech signal as an example of a stochastic pattern and use statistical pattern recognition techniques to produce this word sequence hypothesis.

ASR is mostly defined as solving a maximum a posteriori problem, in which, for a incoming sequence of acoustic vectors A, a sequence of words W must be found such that

$$P(W|A) \tag{1}$$

is optimized, usually under additional constraints imposed by a grammar. Under the general assumption that the Bayesian rule applies, we obtain $P(W|A) = P(A|W)P(W)/P(A)$, and so ($P(A)$ being fixed), the word sequence we are looking for is given by

$$\arg\max_W P(A|W)P(W). \tag{2}$$

The first factor $P(A|W)$ denotes the probability of observing a sequence of acoustic vectors given the word sequence (referred to as acoustic model (AM)), while the second factor $P(W)$ denotes the probability of the word sequence itself (language model (LM)). For a commercial dictation system, the AM is usually trained using an acoustic training database of 50–150 h of speech, while the LM may require a text corpus containing 100–1000 million words. The algorithm that actually performs the optimization of $P(W|A)$ is based on a pattern recognition approach, and is often implemented by using dynamic programming techniques.

The sequence of acoustic vectors A is the result of a properly chosen feature extraction (FE) algorithm. Two properties of the FE are relevant for the discussion here. Firstly, the FE produces a sequence of acoustic feature vectors or 'frames', typically 100 a second. Due to the applied windowing, and the use of delta and delta–delta features, the feature vector represents a sort of fingerprint of the speech spectrum over about 60–70 ms.

Secondly, for regular speech recognition tasks, the FE is designed to normalize for a number of effects that are irrelevant for the decoding into words. These effects include noise, certain channel

effects, line echoes, microphone characteristics, but also speaker-dependent characteristics such as the vocal tract length, and the specific embedding in the entire acoustic space. Speech features that are commonly left out or at least neglected by the FE are the pitch and speaking rate. (Evidently, the pitch is not left out in a number of ASR systems designed to recognize tonal languages such as Chinese, but even in such systems, the pitch is dealt with locally and often not as part of an independent prosodic component on or beyond word level.) 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 in which an utterance is produced, with high or low pitch, fast or slow, angry or sad, these are all irrelevant factors as seen from the 'ASR as a typewriter' point of view.

Another aspect of the most common ASR recognizers, which is of ultimate importance for its modeling power, is the use of hidden Markov models (HMMs). HMMs are used to model the acoustic properties of the recognition tokens, which may be entire words, sub-word units or combinations of words. In the HMM framework, speech is modeled as a two-step probabilistic process (Rabiner and Huang, 1993; Makhoul and Schartz, 1994). In the first step, speech is modeled as a sequence of acoustic states. In the second stage, the acoustic events associated with these states are modeled as probability density functions on the feature space (the precise implementation not being of importance for this paper). Together, these two steps represent the probabilistic modeling of speech in most ASR systems. The output of the ASR system, a list of N-best hypotheses or a word lattice is the result of competition between the AM and the LM. The issue here is, that in the standard HMM training, all HMM states will 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 usually correspond to small time scale events, on segmental or maybe syllable level. The modeling of prosodic patterns associated with emotion, however, only makes sense on a larger time scale spanning a time domain of at least a word. This implies that, in a conventional recognizer, it is feasible to train separate 'variants' for words spoken with different emotions. These variants are considered as 'pronunciation variants' of that particular word. The price to pay is a larger lexicon. It is less likely that such a prosodic/emotion tagging can be performed using speech segments on sub-word level (which would assume something like anger-specific or joy-specific pronunciation of individual sub-word units). The underlying assumption is that the set of acoustic tokens of a particular word spoken with a particular emotion is coherent enough to be described by probability density functions related to a particular sequence of HMM states. Since the acoustic realization of the emotion itself is not directly anchored to the word structure, this coherence will in general be not very strong.

These arguments show that the recognition of emotional speech cannot easily be dealt with in the same way as is usually done in the case of vocal tract length normalization (sometimes called 'gender normalization'), where the optimal choice for the vocal tract length parameter is based on optimization of the likelihood of the utterance given different alignments between the utterance and the AMs. One has been looking for methods to deal with emotions by other means, not using the ASR paradigm (i.e. recognizing tokens on the basis of a frame-by-frame input of spectral parameters), but based on a more direct classification approach with purely prosodic information of the entire utterance as input. This classifier is often modeled as a neural net or as a separate HMM. The next step is to combine information from the prosodic classifier with the word graph from the regular ASR system. A number of studies show the effective combination of prosody and ASR, particularly in small recognition tasks (dialogues). In the next section these will be discussed briefly.

3. Prosodic features and ASR

The attempts done so far to integrate prosodic information and speech recognition have been unsuccessful for improving transcription accuracy.

However, more recently, prosodic information has been incorporated into a variety of ASR-related tasks, such as identifying speech acts, locating focus, improving rejection accuracy (confidence modeling), locating disfluencies, identifying user corrections in dialogue, topic segmentation, and labeling emotions. All these research areas are critically dependent on the collection of appropriate corpora and the development of appropriate prosodic annotation systems.

Evidently, the 'linguistic content' of an utterance goes beyond its 'text' content. 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.

For the integration of supra-segmental information (e.g. 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'.

In the front end method, the FE includes a pitch detection algorithm. The pitch feature is regarded as a separate stream and used to create separate 'acoustic models' (using e.g. a pitch/delta-pitch codebook of Gaussians). The AM used in the 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 a tone language such as Chinese with about 10–30% reduction in syllable error rate. In such an approach, the 'tones' are to be transcribed in the lexicon on the syllable level.

The second, back end method is to include the pitch determination in the FE, but to use the pitch information only for rescoring the *N*-best list or word lattice. In this way, one may use pitch, word stress or other supra-segmental or lexical information to improve the recognition score (see e.g. Streefkerk et al., 1998). Since the back end method can also be used when pitch contour encodes focus, this method is particularly useful for improvement of recognition results in dialogues.

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

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

has a number of very different semantic interpretations, depending on the pitch contour. The use of prosodic information can also be of ultimate importance for studying the dialogues of players talking within a limited discourse domain. In a man–machine dialogue in an information retrieval system, the prosody of the sentence facilitates to implicitly or explicitly fill in the empty or low-confidence slots in the query that is used to retrieve the information. This might be quite complicated: also for a relatively simple task as DARPA's Air Travel Information System (ATIS Technical Report, 1995), the dialogue system needs to go beyond a simple utterance-based keyword spotting scheme to get the meaning of a particular utterance and to react appropriately. In most cases, a correct interpretation of query slots requires knowledge of the dialogue history.

There is an interaction between ASR performance and prosodic properties of the utterance. ASR errors can sometimes, 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 through the FE and affect the acoustic modeling and the test. It is well known that the ASR performance depends on the level of formality and speaking style (Weintraub et al., 1996; Oviatt, 1998). To obtain an optimal ASR result, the ASR test conditions should be 'statistically similar' to the ASR training conditions, and so variations in speaking style and speaking rate a priori have a negative impact on the ASR performance. To speak slower than normal is usually less worse than speaking faster than normal. It is a common effect that customers of a voice operated information system or IVR system tend to hyper-articulate when they cannot get through the dialogue, which is usually a bad strategy to get better recognized (see e.g. Soltau and Waibel, 1998). But prosody can be also used in a positive way. Recently more progress has been claimed in the relation between ASR performance and prosodic properties of utterances (e.g. Hirschberg et al., 2000; Litman et al., 2000; Hirschberg, 1999; Ostendorf et al., 1993; Shriberg et al., 1998). Nöth et al. (1999) show that the integration

of prosodic information might greatly improve the processing speed of the word search. Prosody is capable of reranking the ASR hypothesis such as to distinguish the correctly recognized utterances from incorrectly recognized ones (Veilleux, 1994; Hirose, 1997; Hirschberg et al., 2000). Litman et al. (2000) claim 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 clear whether these prosodic features directly hamper the ASR search (and therefore trivially correlate with word recognition errors) or whether they are more *indirectly* associated with properties in the speech signal that deteriorate ASR performance.

4. Emotion and ASR, affective computing

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 authors 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. (Throughout the remainder of this paper, we will stick to the word emotion, though.)

Quite some research effort is now being put into a field that is called 'affective computing' (Picard, 1997; Affective Computing, 2000). The goal in affective computing is to design ASR and TTS related algorithms (e.g. agents) that understand and respond to human emotions. The commonly applied approach is to start with a database with 'emotional speech' (mostly produced by actors, but recently one attempts to collect corpora with genuine emotional speech, Campbell, 2000). These databases are annotated with emotion tags by a panel of listeners (see e.g. Kienast et al., 1999; Slaney and McRoberts, 1998; Amir and Ron, 1998; Koike et al., 1998). 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 classification step often involves classical techniques closely related to ASR: Gaussian modeling, vector quantization (VQ), artificial neural networks (ANN), and expert networks (cf. Li and Zhao, 1998). In the third step, the resulting parameter estimates are verified and adapted by using a speech synthesis tool, followed by a formal human classification test of the synthesized emotional speech, or by direct integration of the outcome of the emotion classifier into ASR.

In the context of ASR/NLP, an appropriate way to deal with emotion is to deal with an utterance on the following levels:

1. text (segmental) level; this level is accessed by the classical ASR;
2. 'semantic' level, in which the word hypothesis is linked to a 'meaning' (and appropriate action) (NLU, as a part of NLP);
3. prosodic (supra-segmental) level: pitch, volume, pausing, phrasing, speaking rate;
4. emotion level: neutral, sadness, happiness, anger, etc.;
5. 'functional' level: directive, question, approval, attention, prohibition, impatience, frustration, etc.

Paraphrasing Manning (2000), the study of spoken language use deals with the probability distribution

$$P(A, W, T, M),$$

in which A, W, T and M denote the acoustic signals, the word sequences, the syntactic tree structures, and the meanings, respectively. In classical ASR, people look at $P(A, W)$, with the rest of the structure ignored. NLP deals primarily with relations between W, T and M. Language generation is related to $P(W, M)$. Prosody (and certainly emotion) in the speech signal involves A, W and M.

The combination of prosody/emotion and ASR is of particular interest when it comes to understanding and dialogue aspects. Having decoded the speech signal into a (hypothesized) sequence of words, a traditional speech understanding system employs a sentence parser to cast the word

sequence into a syntax structure to allow inference of meaning. Most parsing algorithms focus on syntactic structure first, rather than meaning (understanding), so the coupling is not very tight. At the present time, automatic understanding of speech is limited to determining a specific action based on the speech input. So, in order to use emotion in a dialogue system, it is of interest to see how emotion can help to interpret specific terms, spoken in isolation or embedded in natural sentences, to specify an intended action. We have seen that prosodic information can help in disambiguating utterances and reranking hypotheses; the use of an *emotional* component in the prosody might further complicate the correct interpretation of the utterance and thereby the generation of an appropriate response.

According to many studies (e.g. Picard, 1997; Mozziconacci and Hermes, 1998; Juang and Furui, 2000; Petrushin, 2000; Kang et al., 2000), pitch is the most relevant acoustic parameter for the detection of emotion, followed by energy, duration and speaking rate. In Kienast et al. (1999), the emotions anger, fear, sadness, anxiety and happiness were studied in terms of their prosodic acoustic (and articulatory) correlates. It was found that in a number of cases speaking rate, segment duration and accuracy of articulation are useful parameters to determine the emotion state of the speaker. For example, sadness was clearly shown to correlate with slow speech, while fear was found to correlate with a higher speaking rate than average. 'Anxious' utterances show segments that are shorter than average, with exception of voiceless plosives. Also in (Murray and Arnott, 1993), relations were shown between the emotion state and the duration of vowels and consonants. But in nearly all studies pitch and energy are the most commonly applied features to distinguish and classify emotion state (Murray and Arnott, 1993), or anyway to convey supra-textual information. In Slaney and McRoberts, 1998, a study was conducted to automatically classify an utterance (spoken by a parent to a young infant) into three classes: approval, attention and prohibition. Compared to a system that is to detect emotion states, this classification looks like an easy task, but it appears far from trivial to obtain a reasonable performance. Based on pitch slope, mean pitch and mean delta pitch, measured globally on the entire utterance, the results were close to 55% 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% up to 92% (based on 30–50 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 has been claimed (Zetterholm, 1998).

In an interesting study, Petrushin (2000) compares four different classification strategies to recognize emotion states from the speech input. The study aims at the distinction of five emotions happiness, anger, sadness, fear, and a default 'normal' state. A corpus of emotion data (telephone quality) has been made using utterances from non-professional actors. The FE produced parameter vectors containing $F0$, vocal energy, speaking rate, the first three formants and corresponding bandwidths. Of all these parameters the mean, standard deviation, minimum, maximum and range were evaluated; furthermore, the augmented parameter vector contained the slope of the pitch. The features with highest discriminative power were selected for further processing. The used classifiers were a K-nearest neighbor classifier, neural networks, ensembles of neural networks, and expert networks. Across all classification methods, fear was one of the emotions that were poorly recognized (below 50%); sadness and anger however were recognized with a performance of 70–80%, which, compared to other studies, is a good result.

Nogueiras et al. (2001) report on recognition results using standard HMM techniques. The database they used was taken from the Spanish part of the INTERFACE Emotional Speech Synthesis Database. The speech corpus, produced by a female and a male professional actor, contained an equal number of examples of six emotional styles: anger, disgust, fear, joy, sadness, surprise, and a neutral style. The classifier training was based on

consensus labelling. The confusion matrix of a subjective evaluation showed that the six emotions perceptually felt apart in two major groups: fear, disgust and sadness on one hand, and surprise, joy and anger on the other. The authors use a measure of 'harshness' based on peaks in the autocorrelation function. Single state HMMs were used to represent the probability distribution of all acoustic features (based on pitch, the logarithm of the pitch, the harshness, the energy, the derivatives, and syllabic versions of these), and the emotion tag was hypothesized by an utterance-based maximum likelihood criterion after aligning the input utterance with the emotion-specific HMM model. In an unbiased test set containing 555 utterances, the emotions (including neutral) were recognized with an accuracy of 70% or more in the speaker-independent mode; in the speaker-dependent mode, the accuracy was about 10% higher. It was further shown that about one-sixth of the 'joy' utterances were actually classified as 'surprise', and anger and joy appear less well distinguishable than other emotion pairs, findings which are in line with the human confusion results.

4.1. Synthesis

As already observed earlier, speech synthesis is a technique often used to study the relation between acoustic features and emotion percepts (see e.g. Part 3 on synthesis in Cowie et al., 2000, Proceedings of the ISCA Workshop on Speech and Emotion). Many of the emotion studies use in fact speech synthesis (e.g. Montero et al., 1998; Rank and Pirker, 1998; Whiteside, 1998; Iida et al., 1998; Mizuno and Nakajima, 1998). To simulate emotion states in synthesis, one usually modifies pitch, segment duration and phrasing parameters to create the desired emotion effect. Speech synthesis modules have shown to be a useful tool to study the impact of supra-segmental features on the perception of emotion. The drawback is that the test paradigm in such a setting is quite limited, and it does not cover real-life situations. From an acoustic recording, however, not all interesting acoustic features can easily be accessed; speaking rate is an example of such a feature. Moreover, the recordings will show a larger variation due to the less controlled speech production. For example, the speaking rate correlates with many more speech and speaker characteristics, e.g. with articulatory sloppiness and non-nativeness of the speaker.

Some studies use synthesized emotional speech with a speech synthesizer using parameters such as $F0$, duration and amplitude, but also voice quality parameters, spectral energy distribution, harmonics-to-noise ratio, and articulatory precision. An example of such a study is presented by Rank and Pirker (1998). They focus at the four emotions anger, sadness, fear and disgust. They conclude that sadness is the most 'distinctive' emotion, i.e. the easiest to distinguish from the other three, compared to the other emotions. Whiteside (1998) 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. Their database contained two speakers only—which is too small to draw conclusions about generalization across speakers.

In a number of cases, synthesis model parameters are also based on rules derived from a database with speech with 'emotional prosody'. For Spanish, Montero et al. (1998) implemented a rule-based simulation of three primary emotions into a 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 was about 60–70%. The same technique was applied for Japanese (Iida et al., 1998), in an attempt to improve the expression of the three emotions joy, anger and sadness by using CHATR, the concatenative speech synthesis system developed at ATR. A perceptual experiment was conducted using stimuli synthesized on the basis of analyses on each emotion corpus. $F0$ and duration showed significant differences among emotion types. They showed that mean funda-

mental frequency was lowest for sadness and highest for happiness/joy. Duration per phone for sadness was longest and for anger was shortest. The authors also looked at pauses, but the only significant finding was that pauses were longer in the 'sad' corpus than they were in the other corpora.

4.2. Influence of culture

It may be difficult to identify the emotion of a speaker from a different culture (Shigeno, 1998; Koike et al., 1998; Scherer, 2000). Shigeno (1998) additionally 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 the auditory mode.

4.3. Language dependencies

Emotion patterns may be language dependent (Koike et al., 1998). 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%. Their experimental results suggest that surprise and joy may depend more on semantics, rather than on prosody.

4.4. Linear–non-linear features

Zhou et al. (1998) take another position. Rather than studying the effect of emotion in general, they investigate the effect of a stressful situation on the acoustic speech characteristics. Stressful or highly emotional modes usually deteriorate the performance of a speech recognition system. To address this effect, they study a number of linear and non-linear features and processing methods for the classification of what the authors call stressed speech. The linear features include properties of pitch, duration, energy, and parameters related to the glottal source. The non-linear 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

$$\mathrm{TEO}(x[n]) = x[n]x[n] - x[n+1]x[n-1],$$

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 probability density function has been modeled to match the training vectors—these training vectors were sequences of measurements of the individual features over time. The tests focus 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 equal-error rate based detection clearly indicate that pitch is the best of the five 'linear' features for stress classification (result about 88%); the non-linear TEO-based feature, however, outperforms pitch by about 5%. 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.)

The speech material used by Pereira and Watson (1998) 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.

4.5. Short-term–long-term features

As we could expect, long-term features seem to outperform short-term features (Li and Zhao, 1998). 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 (VQ, ANN 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.

Amir and Ron (1998) discusses a method in which an 'emotion 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 emotion index. In this preliminary study, the authors report success rates of about 50% for 5 emotions (acoustic measurements based on 24 speakers).

5. Discussion and conclusion

In general, the recognition of emotion is not straightforward. Acoustically, emotions overlap and appear in various degrees. Many studies, but not all, support a quite clear distinction between the three 'emotion groups' neutral–sadness, anger–fear and happiness–surprise. Within these emotion groups, the separation appears much more difficult. 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. The acoustic realization of specific emotions seems to be speaker dependent to a large extent, and some cross-language studies indicate that the acoustic realizations of emotions are language dependent.

In general, the most useful phonetic feature for utterance-based emotion decoding is pitch (including derivatives, etc.) followed by energy. One study (Oviatt, 1998) defines a non-linear energy-related feature outperforming pitch in a down-scaled stress detection task. Straightforward Gaussian modeling was shown to be an adequate method to distinguish emotion classes in a space spanned by the following phonetic parameters: pitch, pitch range, average pitch, all measured on the speech part of the utterance (i.e. after removing pauses from the utterances). Almost all studies show that, given a particular speaker, the pitch mean is lowest for sad speech and highest for joy/happiness, and that speaking rate is lowest for sad speech.

The validation of an automatic emotion recognition system is based on subjective judgments from a panel. A number of studies discuss the

difficulty to define an objective scale for subjective phenomena, especially across speakers. That is one reason why an analysis may perform well in a speaker-dependent mode, and worse for all speakers simultaneously (irrespective of the positive effect of speaker adaptive training). This effect seems 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, a form of consensus labeling has been used during training.

Prosodic information is of limited interest for improving the ASR (word) accuracy, but has proven to be very useful in a limited discourse domain. Prosody, and to a smaller extent emotion, is of importance for the semantic and pragmatic disambiguation of certain utterances in dialogues (see e.g. Hirose, 1997; Alter et al., 2000 and this issue). The classification of moods such as frustration and impatience can improve the adequacy of the dialogue. The role of these 'pragmatic factors' or 'assisting factors' is recognized in a number of recent studies (see e.g. Cowie, 2000; Campbell, 2000; Iida et al., 2000).

In most emotion 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 emotion tags could be stronger. As Huber et al. (2000) point out, the main problem in the task of classification of emotion in speech is the need of realistic speech data, such as 'angry people in real situations'.

We conclude with two examples of emotion research in which the recognition of emotion is studied in a larger context. Vyzas et al. (1999) worked on emotion recognition by studying the physiological changes that occurred in an actor who intentionally induced eight different emotion states. These changes were detected with a number of sensors measuring blood volume pressure, skin conductivity (GSR), respiration and a few more activities. The study included extracting and analyzing useful features from the physiological signals of each emotion state, with the intention of developing algorithms that can discriminate between these emotion states. The different states studied in this experiment were: 'no emotion', 'anger', 'hate', 'grief', 'platonic love', 'romantic love', 'joy' and 'reverence', recorded over 3-minute periods each. Several features were extracted from each signal, including the mean and variance. Each emotion was therefore characterized by a set of values-features ranging from 24 to 40, depending on the features included in the analysis. The authors carried out a classification in this space, on a subspace of it, and in one of reduced dimensionality, produced by the Fisher Projection algorithm. Gaussian probability distributions were then fitted to the data. Using these features, classifiers reached an 80% success rate when discriminating among all eight emotions (for one actor).

Our second and final example is a multimodal study presented by Kitazoe et al. (2000). They attempt to integrate both voice and facial expressions. For the speech input, the pitch, energy, and the derivatives were input for an HMM. The facial expressions were represented by black and white pictures of the face and by thermal images which were input for a separate neural network. Although the stimulus material was limited, the authors claim that a combination of modes yields a performance of the emotion state that was significantly higher than the performance obtained without the facial input. From this perspective, the speech signal itself—especially without the text tier—is just a poor channel to detect emotions.

The site http://emotion.salk.edu/Emotion/EmoRes/CompAI/CompAI.html presents an interesting overview of several other studies on (the recognition of) emotion, mostly in the context of artificial intelligence and computational models.

References

Affective Computing, 2000. See e.g. http://affect.media.mit.edu/AC_research.

Alter, K., Rank, E., Kotz, S.A., Toepel, U., Besson, M., Schirmer, A., Friederici, A.D., 2000. Accentuation and emotions – two different systems? In: Proc. ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework, Newcastle, N. Ireland, 5–7 September 2000, Belfast, September 2000, pp. 138–142.

Amir, N., Ron, S., 1998. Towards an automatic classification of emotions in speech. In: Proc. ICSLP 1998, pp. 555–558.

ATIS Technical Report, 1995. In: Proc. ARPA Spoken Language Systems Technology Workshop, Austin, Texas, pp. 241–280.

Campbell, N., 2000. Databases of emotional speech. In: Proc. ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework, Newcastle, N. Ireland, 5–7 September 2000, Belfast, September 2000, pp. 34–39.

Chu, M., Peng, H., 2001. An objective measure for estimating MOS for synthesized speech. In: Proc. Eurospeech 2001, Aalborg, Denmark, pp. 2087–2090.

Cowie, R., 2000. Describing the emotional states expressed in speech. In: Proc. ISCA ITRW at Newcastle, N. Ireland, 5–7 September 2000, Belfast, pp. 11–18.

Cowie, R., Douglas-Cowie, E., 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Proc. ICSLP 1996, pp. 1989–1992.

Cowie, R., Douglas-Cowie, E., Schröder, M., 2000. Speech and emotion: developing a conceptual framework. In: Proc. ISCA ITRW at Newcastle, N. Ireland, 5–7 September 2000, Belfast, pp. 151–188.

Fernandez, R., Picard, R.W., 2000. Modeling driver's speech under stress. In: Proc. ISCA Workshop on Speech and Emotion, Newcastle, September 2000, pp. 219–224.

Frick, R.W., 1985. Communicating emotion: The role of prosodic features. Psychol. Bull. 97, 412–429.

Hirose, K., 1997. Disambiguating recognition results by prosodic features. In: Sagisaka, Y., Campbell, N., Higuchi, N. (Eds.), Computing Prosody: Computational Models for Processing Spontaneous Speech. Springer Verlag, Berlin, pp. 327–342.

Hirose, K., Minematsu, N., Kawanami, H., 2000. Analytical and perceptual study on the role of acoustic features in realizing emotional speech. In: Proc. ICSLP 2000, pp. 369–372.

Hirschberg, J., 1999. Communication and prosody: functional aspects of prosody. In: Proc. ESCA Workshop Dialogue and Prosody, pp. 7–15.

Hirschberg, J., Litman, D., Swerts, M., 2000. Prosodic cues to recognition errors. Paper presented at ASRU 99, Keystone, USA, December 1999.

Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann, H., 2000. Recognition of emotion in a realistic dialogue scenario. In: Proc. ICSLP 2000, pp. 665–668.

Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M., 1998. Acoustic nature and perceptual testing of corpora of emotional speech. In: Proc. ICSLP 1998, pp. 1559–1562.

Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M., 2000. A speech synthesis system with emotion for assisting communication. In: Proc. ISCA ITRW at Newcastle, N. Ireland, 5–7 September 2000, Belfast, pp. 167–173.

Juang, B.-H., Furui, S., 2000. Automatic recognition and understanding of spoken language – a first step towards natural human–machine communication. Proc. IEEE 88 (8), 1142–1165.

Kang, B.-S., Han, C.-H., Lee, S.-T., Youn, D.-H., Lee, C., 2000. Speaker dependent emotion recognition using speech signals. In: Proc. ICSLP 2000, pp. 383–386.

Kienast, M., Paeschke, A., Sendlmeier, W., 1999. Articulatory reduction in emotional speech. In: Proc. Eurospeech 1999, pp. 117–120.

Kitazoe, T., Kim, S.-I., Yoshitomi, Y., Ikeda, T., 2000. Recognition of emotional states using voice, face image and thermal image of face. In: Proc. ICSLP 2000, pp. 653–656.

Klein, J., 1999. Computer response to user frustration. ftp://white-chapel.media.mit.edu/pub/tech-reports/TR-480-ABSTRACT.html.

Koike, K., Suzuki, H., Saito, H., 1998. Prosodic parameters in emotional speech. In: Proc. ICSLP 1998, pp. 679–682.

Li, Y., Zhao, Y., 1998. Recognizing emotions in speech using short-term and long-term features. In: Proc. ICSLP 1998, pp. 2255–2258.

Litman, D., Hirschberg, J., Swerts, M., 2000. Predicting automatic speech recognition performance using prosodic cues. In: Proc. NAACL, Seattle, May 2000.

Makhoul, J., Schartz, J., 1994. State of the art in continuous speech recognition. In: Roe, D., Wilpon, J. (Eds.), Voice Communication Between Humans and Machines. National Academy Press, pp. 165–188.

Manning C., 2000. Probabilistic models in computational linguistics. Available: http://www.ima.umn.edu/talks/workshops/10-30-11-3.2000/manning/ima2000.pdf.

Mizuno, O., Nakajima, S., 1998. A new synthetic speech/sound control language. In: Proc. ICSLP 1998, pp. 2007–2010.

Montero, J.M., Gutierrez-Arriola, J.M., Palazuelos, S., Enriquez, E., Aguilera, S., Pardo, J.M., 1998. Emotional speech synthesis: from speech database to TTS. In: Proc. ICSLP 1998, pp. 923–926.

Mozziconacci, S., 2000. The expression of emotion considered in the framework of an intonation model. In: Proc. ISCA ITRW at Newcastle, N. Ireland, September 2000, pp. 45–52.

Mozziconacci, S., Hermes, D., 1998. Study of intonation patterns in speech expressing emotion or attitude: production and perception. IPO Annual Progress Report, IPO, Eindhoven.

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

Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J., 2001. Speech emotion recognition using hidden Markov models. In: Proc. Eurospeech 2001, Aalborg, Denmark.

Nöth, E., Batliner, A., Warnke, V., Haas, J., Boros, M., Buckow, J., Huber, R., Gallwitz, F., Nutt, M., Niemann, H., 1999. On the use of prosody in automatic dialogue understanding. In: Proc. ESCA Workshop Dialogue and Prosody, Eindhoven, The Netherlands, September 1999, pp. 25–34.

Ostendorf, M., Wightman, C., Veilleux, N., 1993. Parse scoring with prosodic information: a synthesis/analysis approach. Comput. Speech Lang. 7 (3), 193–210.

Oviatt, S.L., 1998. The CHAM model of hyperarticulate adaptation during human–computer error resolution. In: Proc. ICSLP 1998, pp. 2311–2314.

Pereira, C., Watson, C., 1998. Some acoustic characteristics of emotion. In: Proc. ICSLP 1998, pp. 927–930.

Petrushin, V.A., 2000. Emotion recognition in speech signal: experimental study, development, and application. In: Proc. ICSLP 2000, pp. 222–225.

Picard, R.W., 1997. Affective Computing. MIT Press, Cambridge, MA.

Rabiner, L., Huang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ.

Rank, E., Pirker, H., 1998. Generating emotional speech with a concatenative synthesizer. In: Proc. ICSLP 1998, pp. 671–674.

Scherer, K., 1981. Speech and emotional states. In: Darby, J. (Ed.), Speech Evaluation in Psychiatry. Grune and Stratton, New York, pp. 189–220.

Scherer, K., 2000. A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In: Proc. ICSLP 2000, pp. 379–382.

Shigeno, S., 1998. Cultural similarities and differences in the recognition of audio-visual speech stimuli. In: Proc. ICSLP 1998, pp. 281–284.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, M., Meteer, M., Van-Ess-Dykema, C., 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? Lang. Speech 41, 439–447.

Slaney, M., McRoberts, G., 1998. Baby ears: a recognition system for affective vocalizations. In: Proc. ICASSP, Seattle, WA (on cdrom). Also available via: http://rvl4.ecn.purdue.edu/~malcolm/interval/1997-063/.

Soltau, H., Waibel, A., 1998. On the influence of hyperarticulated speech on recognition performance. In: Proc. ICSLP 1998, pp. 229–232.

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. In: Proc. ICSLP 1998, pp. 683–686.

Van Bezooijen, R., 1984. The Characteristics and Recognizability of Vocal Expressions of Emotion. Foris, Dordrecht, The Netherlands.

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

Vyzas, E., Minka, T., Healey, J., 1999. http://www.media.mit.edu/affect/AC_research/projects/emotion_recognition.html.

Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A., 1996. Effect of speaking style on LVCSR performance. In: Proc. ICSLP 1996, pp. 1457–1460.

Whiteside, S.P., 1998. Simulated emotions: an acoustic study of voice and perturbation measures. In: Proc. ICSLP 1998, pp. 699–703.

Yang, L., 2000. The expression and recognition of emotions through prosody. In: Proc. ICSLP 2000, pp. 74–77.

Zetterholm, E., 1998. Prosody and voice quality in the expression of emotions. In: SST Proc. of the Seventh Australian International Conference on Speech Science and Technology, Sydney, pp. 109–113.

Zhao, L., Lu, W., Jiang, Y., Wu, Z., 2000. A study on emotional feature recognition in speech. In: Proc. ICSLP 2000, pp. 961–964.

Zhou, G., Hansen, J.H.L., Kaiser, J.F., 1998. Linear and nonlinear speech feature analysis for stress classification. In: Proc. ICSLP 1998, pp. 883–886.