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

Speech Communication 41 (2003) 603–623

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speech emotion recognition using hidden Markov models

Tin Lay Nwe [a,*], Say Wei Foo [b], Liyanage C. De Silva [a]

[a] Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Singapore 117576, Singapore

[b] School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore

Received 28 June 2003; received in revised form 28 June 2003; accepted 30 June 2003

Abstract

In emotion classification of speech signals, the popular features employed are statistics of fundamental frequency, energy contour, duration of silence and voice quality. However, the performance of systems employing these features degrades substantially when more than two categories of emotion are to be classified. In this paper, a text independent method of emotion classification of speech is proposed. The proposed method makes use of short time log frequency power coefficients (LFPC) to represent the speech signals and a discrete hidden Markov model (HMM) as the classifier. The emotions are classified into six categories. The category labels used are, the archetypal emotions of Anger, Disgust, Fear, Joy, Sadness and Surprise. A database consisting of 60 emotional utterances, each from twelve speakers is constructed and used to train and test the proposed system. Performance of the LFPC feature parameters is compared with that of the linear prediction Cepstral coefficients (LPCC) and mel-frequency Cepstral coefficients (MFCC) feature parameters commonly used in speech recognition systems. Results show that the proposed system yields an average accuracy of 78% and the best accuracy of 96% in the classification of six emotions. This is beyond the 17% chances by a random hit for a sample set of 6 categories. Results also reveal that LFPC is a better choice as feature parameters for emotion classification than the traditional feature parameters.

Keywords: Recognition of emotion; Emotional speech; Log frequency power coefficients; Hidden Markov model; Human communication

1. Introduction

There are many motivations in identifying the emotional state of speakers. In human–machine interaction, the machine can be made to produce more appropriate responses if the state of emotion of the person can be accurately identified. Most state-of-the-art automatic speech recognition systems resort to natural language understanding to improve the accuracy of recognition of the spoken words. Such language understanding can be further improved if an emotional state of the speaker can be extracted, and this in turn will enhance the accuracy of the system. In general, translation is required to carry out communications using different languages. Current automatic translation algorithms focus mainly on the semantic part of the speech. It would provide the communicating parties an additional useful information if an emotional state of the speaker can also be identified and presented, especially in non-face-to-face

* Corresponding author. Tel.: +65-67904848; fax: +65-67933318.

E-mail addresses: engp8469@nus.edu.sg (T.L. Nwe), eswfoo@ntu.edu.sg (S.W. Foo).

0167-6393/$ - see front matter
doi:10.1016/S0167-6393(03)00099-2

situations. Other applications of automatic emotion recognition systems include, tutoring, alerting, and entertainment (Cowie et al., 2001).

Before delving into the details of automatic emotion recognition, it is appropriate to have some understanding of psychological, biological, and linguistic aspects of emotion (Cowie et al., 2001; Cornelius, 1996; Oatley and Johnson-Laird, 1995; Plutchik, 1994; Scherer, 1986a; Scherer, 1984; Otaley and Jenkis, 1996; Arnold, 1960; Lazarus, 1991; Fox, 1992; Darwin, 1965; Ekman and Friesen, 1975; Schubiger, 1958; O'Connor and Arnold, 1973; Williams and Stevens, 1981; Cowan, 1936; Fairbanks and Pronovost, 1939; Lynch, 1934; Frick, 1985; Murray and Arnott, 1993; Crystal, 1969; Crystal, 1975; Fonagy, 1978a,b; Fonagy and Magdics, 1963; Davitz, 1964; Williams and Stevens, 1969; Van Bezooijen, 1984; Kotlyar and Mozorov, 1976; Muller, 1960; Oster and Risberg, 1986; McGilloway et al., 1995; Trojan, 1952; Havrdova and Moravek, 1979; Huttar, 1968; Coleman and Williams, 1979; Kaiser, 1962; Scherer, 1986b; Utsuki and Okamura, 1976; Sulc, 1977; Johnson et al., 1986). From the psychological perspective, of particular interest is the *cause-and-effect* of emotion (Cornelius, 1996; Oatley and Johnson-Laird, 1995; Plutchik, 1994; Scherer, 1986a; Scherer, 1984; Otaley and Jenkis, 1996; Arnold, 1960; Lazarus, 1991). The activation–evaluation space (Cowie et al., 2001) provides a simple approach in understanding and classifying emotions. In a nutshell, it considers the stimulus that excites the emotion, the cognition ability of the agent to appraise the nature of the stimulus and subsequently his/her mental and physical responses to the stimulus. The mental response is in the form of emotional state. The physical response is in the form of fight or flight [1], or as described by Fox (1992), approach or withdrawal. From a biological perspective, Darwin (1965) looked at the emotional and physical responses as distinctive action patterns selected by evolution because of their survival value. Thus, emotional arousal will have an effect on, the heart rate, skin resistivity, temperature, pupillary diameter, and muscle activity, as the agent prepares for fight or flight. As a result, the emotional state is also manifested in spoken words and facial expressions (Ekman and Friesen, 1975).

Emotional states have a definite temporal structure (Otaley and Jenkis, 1996): For example, people with emotional disorders such as, manic depression or pathological anxiety may be in those emotional states for months and years, or one may be in a bad 'mood' for weeks and months, or emotions such as Anger and Joy may be transient in nature and last no longer than a few minutes. Thus, emotion has a broad sense and a narrow sense effect. The broad sense reflects the underlying long-term emotion and the narrow sense refers to the short-term excitation of the mind that prompts people to action. In automatic recognition of emotion, a machine would not distinguish if the emotional state were due to long-term or short-term effect so long as it is reflected in the speech or facial expression.

The output of an automatic emotion recognizer will naturally consist of labels of emotion. The choice of a suitable set of labels is important. Linguists have a large vocabulary of terms of describing emotional states. Schubiger (1958) and O'Connor and Arnold (1973) used 300 labels between the states in their studies. The 'palette theory' (Cowie et al., 2001) suggests that basic categories be identified to serve as primaries and mixing may be done in order to produce other emotions similar to the mixing of primary colors to produce all other colors. The 'primary' emotions that are often used include, Joy, Sadness, Fear, Anger, Surprise and Disgust. They are often referred to as *archetypal emotions*. Although these archetypal emotions cover a rather small part of emotional life, they nevertheless represent the popularly known emotions and are recommended for testing the capabilities of an automatic recognizer. Cognitive theory would argue against

[1] *Fight or Flight* is a physiological/psychological response to a threat. During this automatic, involuntary response, an area of the brain stem will release increased quantity of NOREPINEPHRINE which in turn causes the ADRENAL glands to release more ADRENALINE. This increase in Adrenaline causes faster heart rate, pulse rate, respiration rate. There is also, shunting of the blood to more vital areas, and release of blood sugar, lactic acid and other chemicals, all of which is involved in getting the body ready for fighting the danger (a tiger, a mugger), or running away from the threat. Feelings of dread, fear, impending doom, are common.

equating emotion recognition with assigning category labels. Instead, it would want to recognize the way a person perceives the world or key aspects of it. Perhaps it is true to say that category labels are not a sufficient representation of emotional state, but they are a better way to indicate the output from an automatic emotion recognition system.

It is to be noted that the emotional state of a speaker can be identified from the facial expression (Ekman, 1973; Davis and College, 1975; Scherer and Ekman, 1984), speech (McGilloway et al., 2000; Dellaert et al., 1996; Nicholson et al., 1999), perhaps brainwaves, and other biological features of the speaker. Ultimately, a combination of these features may be the way to achieve high accuracy of recognition. In this paper, the focus is on emotional speech recognition.

The remainder of this paper is structured as follows. In Section 2, a review of the features relevant to emotion of speech is presented and in Section 3, some of the speech emotion recognizers are discussed. This is followed by a description of the corpus of emotional speech and presentation of results of subjective assessment of the emotional content of the speech. Details of the proposed system are presented in Section 5. Experiments to assess the performance of the proposed system are described in Section 6 together with analysis of the results of the experiments. The concluding remarks are presented in Section 7.

2. Characteristics of emotional speech

There are two broad types of information in speech. The semantic part of the speech carries linguistic information insofar that the utterances are made according to the rules of pronunciation of the language. Paralinguistic information, on the other hand, refers to the implicit messages such as the emotional state of the speaker. For speech emotion recognition, the identification of the paralinguistic features that represent the emotional state of the speaker is an important first step.

From the perspective of physiology in the production of speech, Williams and Stevens (1981) stated that the sympathetic nervous system is aroused with the emotions of Anger, Fear or Joy. As a result, heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is correspondingly loud, fast and enunciated with strong high frequency energy. On the other hand, with the arousal of the parasympathetic nervous system, as with Sadness, heart rate and blood pressure decrease and salivation increases, producing speech that is slow and with little high frequency energy. The corresponding effects on speech of such physiological changes thus show up in the overall energy, energy distribution across the frequency spectrum and the frequency and duration of pauses of speech signal.

From the reported findings on features of speech and emotional states (Schubiger, 1958; O'Connor and Arnold, 1973; Williams and Stevens, 1981; Cowan, 1936; Fairbanks and Pronovost, 1939; Lynch, 1934; Frick, 1985; Murray and Arnott, 1993; Crystal, 1969; Crystal, 1975), three broad types of speech variables have been identified as related to the expression of emotional states. These are fundamental frequency (F0) contour, continuous acoustic variables, and voice quality, respectively. Fundamental frequency contour has been used to describe Fundamental frequency variation in terms of geometric patterns. Continuous acoustic variables include magnitude of fundamental frequency, intensity, speaking rate, and distribution of energy across the spectrum. These acoustic variables are also referred to as the augmented prosodic domain. The terms used to describe voice quality are tense, harsh, and breathy. These three broad types of speech variables are somewhat interrelated. For example, the information of fundamental frequency and voice quality are reflected and captured by certain continuous acoustic variables.

A summary of the relationships between six archetypal emotions and the three types of speech parameters mentioned above is presented in Table 1(panels A and B). The data are taken from several sources as indicated in the table. From the data, it is clear that continuous acoustic variables provide reliable indication of the emotions. It also shows that there are contradictory reports on certain variables such as the speaking rate for the Anger

Table 1
Characteristics of specific emotions (panels A and B)

Emotions	Anger	Surprise	Joy
Panel A			
Pitch contour	Angular frequency curve (Fonagy, 1978a), stressed syllables ascend frequently and rhythmically (Fonagy and Magdics, 1963), irregular up and down inflection (Davitz, 1964), level average pitch except for jumps of about a musical fourth or fifth on stressed syllables (Fonagy and Magdics, 1963)	Sudden glide up to a high level within the stressed syllables, then falls to mid-level or lower level in last syllable (Fonagy and Magdics, 1963)	Descending line, melody ascending frequently and at irregular intervals (Fonagy and Magdics, 1963)
Continuous acoustic variables			
Average pitch	Increased in mean (Fonagy, 1978a; Davitz, 1964; Williams and Stevens, 1969; Van Bezooijen, 1984)	–	Increased in mean (Murray and Arnott, 1993; Davitz, 1964; Fonagy, 1978a; Van Bezooijen, 1984; Oster and Risberg, 1986)
Pitch range	Much wider (Fairbanks and Pronovost, 1939; Murray and Arnott, 1993; Williams and Stevens, 1969)	Wide range (Fonagy and Magdics, 1963), median, normal or higher (Oster and Risberg, 1986)	Much wider (Murray and Arnott, 1993; Havrdova and Moravek, 1979; Fonagy and Magdics, 1963)
Intensity	Raised (Davitz, 1964; Williams and Stevens, 1969; Van Bezooijen, 1984; Kotlyar and Mozorov, 1976)	–	Increased (Murray and Arnott, 1993; Van Bezooijen, 1984; Huttar, 1968)
Rate	High rate (Murray and Arnott, 1993; Davitz, 1964; Muller, 1960; Fonagy, 1978b), reduced rate (Oster and Risberg, 1986)	Tempo normal (Oster and Risberg, 1986), tempo restrained (Fonagy and Magdics, 1963)	Increased rate (Davitz, 1964; Coleman and Williams, 1979), slow temp (Van Bezooijen, 1984)
Spectral	High midpoint for average spectrum for nonfricative portions (McGilloway et al., 1995)	–	Increase in high frequency energy (Van Bezooijen, 1984; Kaiser, 1962)
Voice quality	Tense (Fonagy, 1978b), breathy (Murray and Arnott, 1993; Trojan, 1952), heavy chest tone (Murray and Arnott, 1993; Trojan, 1952), blaring (Davitz, 1964)	Breathy (Fonagy and Magdics, 1963)	Tense (Scherer, 1986b), breathy (Murray and Arnott, 1993; Fonagy and Magdics, 1963), blaring tone (Murray and Arnott, 1993; Davitz, 1964)
	Fear	Disgust	Sadness
Panel B			
Pitch contour	Disintegration of pattern and great number of changes in direction of pitch (Fairbanks and Pronovost, 1939)	Wide, downward terminal inflects (Murray and Arnott, 1993)	Downward inflections (Davitz, 1964)
Continuous acoustic variables			
Average pitch	Increase in mean F0 (Fonagy, 1978a; Coleman and Williams, 1979; Utsuki and Okamura, 1976)	Very much lower (Murray and Arnott, 1993)	Below normal mean (Murray and Arnott, 1993; Williams and Stevens, 1969; Coleman and Williams, 1979)
Pitch range	Increase in range F0 (Williams and Stevens, 1969; Utsuki and Okamura, 1976)	Slightly wider (Murray and Arnott, 1993)	Slightly narrower (Murray and Arnott, 1993; Fonagy, 1978a; Williams and Stevens, 1969)

Intensity	Normal	Lower (Murray and Arnott, 1993)	Decreased (Murray and Arnott, 1993; Davitz, 1964; Muller, 1960)
Rate	Increased rate (Kotlyar and Mozorov, 1976, Coleman and Williams, 1979), Reduced rate (Sulc, 1977)	Very much faster (Murray and Arnott, 1993)	Slightly slow (Murray and Arnott, 1993; Fonagy, 1978b; Johnson et al., 1986), long pitch falls (Davitz, 1964)
Spectral	Increase in high-frequency energy		Downward inflections (Davitz, 1964)
Voice quality	Tense (Scherer, 1986b), irregular voicing (Murray and Arnott, 1993)	Grumble chest tone (Murray and Arnott, 1993)	Lax (Scherer, 1986b), resonant (Murray and Arnott, 1993; Davitz, 1964)

emotion. It is also noticed that some speech attributes seem to be associated with general characteristics of emotion, rather than with individual categories. For example, Anger, Fear, Joy and to a certain extent, Surprise has positive activation (approach) and hence have similar characteristics such as much higher average of F0 values and much wider F0 range. On the other hand, emotions such as Disgust, Sadness and to a lesser extent boredom that are associated with negative activation (withdrawal) have lower average of F0 values and narrower fundamental frequency range. The similarity of acoustical features for certain emotions implies that they can easily be mistaken for one another as observed by Cahn (1990). This suggests that grouping of emotions with similar characteristics may improve the system performance.

3. Review of emotional speech classifiers

Although there are a number of systems proposed for emotion recognition based on facial expressions, only a few systems based on speech input are reported in the literature.

ASSESS (McGilloway et al., 2000) is a system that makes use of a few landmarks—peaks and troughs in the profiles of fundamental frequency, intensity and boundaries of pauses and fricative bursts in identifying four archetypal emotions, viz. Fear, Anger, Sadness and Joy. Using *discriminant analysis* to separate samples that belong to different categories, classification rate of a 55% was achieved.

Dellaert et al. (1996) focused on the F0 information for classification. Four emotions, viz. Joy, Sadness, Anger and Fear were considered. It was reported that the most salient features which represent the acoustical correlates of emotion are maximum, minimum and median of the fundamental frequency and the mean positive derivative of the regions where the F0 curve is increasing. Using *K-nearest neighbours* as classifier and majority voting of specialists, the best accuracy achieved in recognition of four emotions was 79.5%.

Nicholson et al. (1999) analysed the speech of radio actors involving eight different emotions. The

emotions chosen were Joy, Teasing, Fear, Sadness, Disgust, Anger, Surprise and Natural. In the study, which was limited to emotion recognition of phonetically balanced words, both prosodic features and phonetic features were investigated. Prosodic features used were speech power and fundamental frequency while phonetic features adopted were *Linear Prediction Coefficients* (LPC) and the Delta LPC parameters. A neural network was used as the classifier. The best accuracy achieved in classification of the eight emotions was 50%.

Cahn (1990) explored improvements to the affective component of synthesized speech using an effect editor program. The input of this program was an acoustical description of emotion and an utterance and the output was synthesized expressive speech. The effects in synthesized speech were achieved by carefully controlling speech parameters such as fundamental frequency, timing, voice quality and articulation. Synthesized speech was generated and selected participants were asked to choose from among six effects, Angry, Disgusted, Glad, Sad, Scared or Surprised. It was reported that, except for Sadness, with a 91% recognition rate, the intended emotions were recognized in approximately 50% of the presentations. The work concluded that Sadness with the most acoustically distinct features, soft, slow, halting speech with minimal high frequency energy, was the most recognizable. Emotions with similar acoustical features, such as Joy and Surprise or Anger and Surprise, were often confused.

The features adopted by most emotion classification research focus on statistics of fundamental frequency, energy contour, duration of silence and voice quality (McGilloway et al., 2000; Dellaert et al., 1996; Nicholson et al., 1999; Cahn, 1990). However, as mentioned above, certain emotions have very similar characteristics based on this set of features. Hence systems based on these features for emotion classification are not able to accurately distinguish more than a couple of emotion categories. This motivates us to search for other acoustic features to identify human emotion in speech. The fundamental frequency is considered as one of the important features in emotion classification (Schubiger, 1958; O'Connor and Arnold, 1973; Williams and Stevens, 1981; Cowan, 1936; Fairbanks and Pronovost, 1939; Lynch, 1934; Frick, 1985; Murray and Arnott, 1993; Crystal, 1969; Crystal, 1975; McGilloway et al., 2000; Dellaert et al., 1996; Nicholson et al., 1999; Cahn, 1990). Furthermore, as stated in (Williams and Stevens, 1981), speech is loud, fast and enunciated with strong high frequency for Anger or Joy emotions. For Sadness, the speaking rate is slow and minimal high frequency energy is observed (Williams and Stevens, 1981; Cahn, 1990). Features that exploit these characteristics will be good indicators for emotion in speech. In this paper, a recognizer is proposed that adopts LFPC as feature parameters of speech to represent energy distribution across the frequency spectrum and a four-stage ergodic HMM is used as the classifier to take care of the effects of speaking rate and variation of F0.

4. Database

4.1. Emotion corpus

An emotion database is specifically designed and set up for text-independent emotion classification studies. The database includes short utterances covering the six archetypal emotions, namely Anger, Disgust, Fear, Joy, Sadness and Surprise. Non-professional speakers are selected to avoid exaggerated expression. A total of six native Burmese language speakers (three males and three females), six native Mandarin language speakers (three males and three females) are employed to generate 720 utterances. Sixty different utterances, ten each for each emotional mode, are recorded for each speaker. The recording is done in a quiet environment using a mouthpiece microphone.

A set of sample sentences translated into the English language is presented in Table 2. Statistics of the durations of the utterances for each of the six emotion categories are given in Table 3. The durations of the utterances for the six categories of emotion are evenly spreaded and the effect of length as a clue for classification is minimal.

In this paper, utterances in Burmese and Mandarin are used due to an immediate availability of

Table 2
Sample sentences of emotion database

Emotion	Sentence
Anger	You are always late (B: Burmese) Don't come here (B) You are not fair to me (M: Mandarin) I am going to kill you (M)
Dislike	I don't want to wear this dress (B) I don't like this color (B) I don't like the way you do it! (M) I don't want to go (M)
Fear	Nobody accompanies me (B) The water is too deep (B) I am going to die (M) He will deteriorate and die (M)
Joy	Hey, you pass the exam (B) Your baby is so cute (B) I have succeeded. I won (M) I struck the first prize of one million (M)
Sadness	I feel so sad (B) My little dog died (B) Your daddy isn't coming back (M) Life is meaningless. I want to die (M)
Surprise	Snake! Snake! (B) Five distinctions! (B) He has woken up (M) Is it real? (M)

native speakers of the languages. It is easier for the speakers to express emotions in their native language than in a foreign language. Since emotional speech is independent of language (Scherer et al., 2001), the language used does not significantly affect the approach taken as can be observed from the results.

4.2. Classification by human subjects

Subjective assessment of the emotional speech corpus by human subjects was carried out. One of the objectives of the subjective classification is to determine the ability of listeners to correctly classify the emotional modes of the utterances. Another objective is to compare the accuracy of classification by the proposed system with human classification performance.

Three subjects of different language background are engaged for the subjective tests. The utterances were played back in random order and the subjects were requested to indicate which one of the six emotional modes was portrayed. The utterances were presented via headphones and repeated two times. The language of the utterances presented to the human subject was neither his mother tongue nor any other language that he has any knowledge to perceive linguistically. Hence the judgment was made based on the perceived emotional content rather than the context of the utterances.

The performance of the human evaluators was summarized in Table 4 and the corresponding confusion matrix is shown in Table 5(panels A and B). Results show that high accuracy is observed in the classification of Anger and Sadness as they have the most acoustically distinct features.

Table 3
Lengths of sample speech utterances for Mandarin and Burmese speakers (Sec)

Anger	Disgust	Fear	Joy	Sadness	Surprise
Burmese					
0.33	0.5	0.51	0.4	0.54	0.31
0.66	1.17	1.28	0.64	1.31	0.84
1.44	1.83	1.86	1.57	1.97	1.31
1.72	2.28	2.45	2.85	2.33	1.85
Mandarin					
0.28	0.43	0.46	0.41	0.51	0.42
0.64	1.49	1.73	1.75	1.25	1.37
1.26	1.99	2.33	2.1	2.43	2.2
1.98	2.68	3.1	2.64	3.04	3.49

Table 4
Average accuracy of human classification (%)

Speaker	Average performance (human)
S1 Burmese (male)	75
S2 Burmese (male)	65
S3 Burmese (male)	61.7
S4 Burmese (female)	76.7
S5 Burmese (female)	60
S6 Burmese (female)	71.7
S1 Mandarin (male)	61.7
S2 Mandarin (male)	58.3
S3 Mandarin (male)	66.7
S4 Mandarin (female)	63.3
S5 Mandarin (female)	66.7
S6 Mandarin (female)	61.7
Mean	65.7

5. The proposed system

5.1. Overview of the system

As mentioned in Sections 2 and 3, the classification based on emotion on certain specific features is not clearly defined. Thus, instead of resorting to measurement of specific features of the speech signals such as fundamental frequency contour to identify the type of emotion, the novel acoustic feature that can distinguish several emotion categories is proposed in this paper.

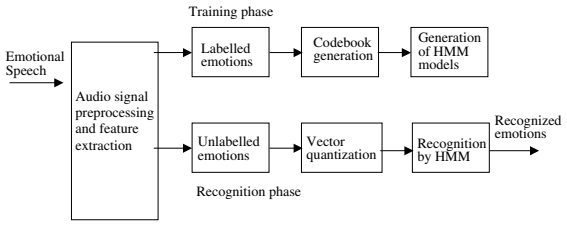

Fig. 1. Block diagram of the proposed system.

The block diagram of the proposed system is shown in Fig. 1. The speech signal is sampled at 22.05 kHz and coded with 16 bits PCM. The signal samples are segmented into frames of 16 ms each with 9 ms overlap between consecutive frames. The typical values of fundamental frequency of speakers ranges from 100 to 200 Hz. The window size of duration 16 ms, covers approximately two periods of fundamental frequency (Cairns and Hansen, 1994). The total number of frames, N, to be processed, depends on the length of the utterance.

For each frame, a feature vector based on normalized LFPC is obtained. For the training session, a universal codebook is constructed using the feature vectors of all utterances reserved for training. HMM models are built for all six emotions individually and for the combined emotions for the reduced set emotion classification.

To reduce the complexity in the classification process, vector quantization is carried out to

Table 5
Human performance confusion matrix for Burmese speakers (panel A) and Mandarin speakers (panel B)

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Panel A						
Anger	***59***			4		5
Disgust	1	***38***	8	3	1	3
Fear		4	***27***	4	7	6
Joy		3	1	***32***	1	6
Sadness		12	19	4	***51***	1
Surprise		3	5	13		***39***
Panel B						
Anger	***58***					1
Disgust	2	***33***	11	12		3
Fear		5	***27***	9	5	19
Joy		16	8	***25***		8
Sadness		3	14	13	***55***	
Surprise		3		1		***29***

convert the feature vector to a single code that represents the best match codebook entry. The resultant sequence of codes for each utterance is then submitted to the HMM classifier. Details of the various stages are presented in the following subsections.

5.2. Selection of feature vectors

It is recognized that the intensity, the spectral distribution, the speaking rate, and the fundamental frequency contour are important features that discriminate the emotional state in speech. Surprise and Anger tend to have the highest intensity, the highest speaking rate and the highest frequency content while Sadness and Disgust have the lowest intensity, the lowest speaking rate and the lowest frequency content.

One possible measure of the emotional content of speech is thus the distribution of the spectral energy across the speech range of frequency. The energy contents of the sub-bands provide essential information on energy distribution and overall intensity. Hence they are selected as representative parameters for the 'emotional' content of speech. By analyzing these data using an HMM recognizer, the effects of the speaking rate and the variation of tone are also taken into consideration.

For speech recognition, LPCC (Atal, 1974) and MFCC (Davis and Mermelstein, 1980) are the popular choices as features representing the phonetic content of speech (Rabiner and Schafer, 1978). However, in this study, it is found that the short time LFPC gives better performance for recognition of speech emotion compared with LPCC and MFCC that are used for speech recognition. A possible reason is that LFPCs are more suitable for the preservation of fundamental frequency information in the lower order filters.

A Log frequency filter bank can be regarded as a model that follows the varying auditory resolving power of the human ear for various frequencies. The filter bank is designed to divide speech signal into 12 frequency bands that match the critical perceptual bands of the human ear. The block diagram of sub-band processing is shown in Fig. 2. The center frequencies f_i and bandwidths b_i

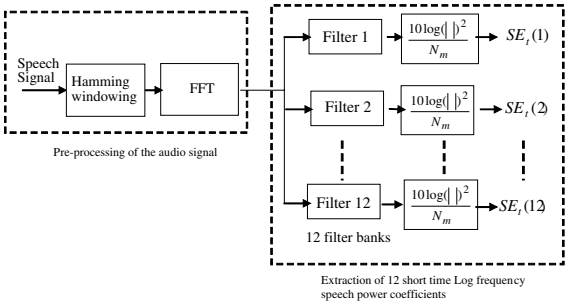

Fig. 2. Block diagram of sub-band processing.

for a set of 12 bandpass filters are derived as follows (Rabiner and Juang, 1993):

$$b_1 = C \tag{1}$$

$$b_i = \alpha b_{i-1}, \quad 2 \leqslant i \leqslant 12 \tag{2}$$

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_1)}{2} \tag{3}$$

where, C is the bandwidth, f_1 is the center frequency of the first filter and α is the logarithmic growth factor. To make use of the information of the fundamental frequency, the frequency of the lower band is set at 100 Hz. Hence $C = 54$ Hz and $f = 127$ Hz. Normal human beings perceive audible sound from 20 Hz to 20 kHz. Although speech signals covering the frequency range from 200 Hz and 3.2 kHz is sufficient for intelligibility (Rabiner and Juang, 1993), useful information for emotional speech recognition may be prevalent beyond the 3.2 kHz. Experiments with different frequency bands up to 7.2 kHz were carried out by changing the value of $\alpha \in [1, 1.4]$. The center frequencies and the bandwidths of the 12 bands for different values of α are given in Table 6.

5.3. Computation of short time LFPC

The samples of each frame are weighted with a Hamming window to reduce spectral leakage. This windowed speech is transformed to the frequency domain using the discrete fourier transform (DFT) algorithm. The spectral components are separated into 12 bands. The DFT responses of the 12 filters are simply the shifted and frequency warped

Table 6
Center frequencies (CF) and bandwidths (BW) of log frequency filter banks for different values of α

Filter	$\alpha = 1.0$		$\alpha = 1.1$		$\alpha = 1.2$		$\alpha = 1.3$		$\alpha = 1.4$	
	CF	BW	CF	BW	CF	BW	CF	BW	CF	BW
1	127	54	127	54	127	54	127	54	127	54
2	181	54	184	59	186	65	189	70	192	75
3	235	54	246	65	258	78	270	91	281	104
4	289	54	315	72	343	93	375	119	406	145
5	343	54	390	79	446	112	511	154	579	202
6	397	54	473	87	569	134	689	201	820	280
7	451	54	564	96	717	161	919	261	1155	389
8	505	54	665	105	894	193	1219	339	1620	541
9	559	54	775	116	1107	232	1609	440	2267	753
10	613	54	897	127	1363	279	2115	573	3167	1046
11	667	54	1031	140	1669	334	2774	744	4417	1454
12	721	54	1178	154	2037	401	3630	968	6154	2021

versions of a rectangular window $W_m(k)$ (Becchetti and Ricotti, 1998).

$$W_m(k) = \begin{cases} 1 & l_m \leqslant k \leqslant h_m \\ 0 & \text{otherwise} \end{cases}, \quad m = 1, 2, \ldots, 12 \quad (4)$$

where k is the DFT domain index, l_m and h_m are the lower and upper edges of mth filter bank. The mth filter bank output is given by:

$$S_t(m) = \sum_{k=f_m-(b_m/2)}^{f_m+(b_m/2)} \left(X_t(k)W_m(k)\right)^2, \quad m = 1, 2, \ldots, 12 \quad (5)$$

where, $X_t(k)$ is the kth spectral component of the windowed signal, t is the frame number, $S_t(m)$ is the output of the mth filter bank, and f_m, b_m is the center frequency and the bandwidth of the mth sub-band respectively.

The parameters $SE_t(m)$ which provide an indication of energy distribution among sub-bands are calculated as follows.

$$SE_t(m) = \frac{10\log_{10}(S_t(m))}{N_m} \quad (6)$$

where, N_m is the number of spectral components in the mth filter bank. For each speech frame, 12 LFPCs are obtained.

5.4. Computation of LPCC and MFCC coefficients

For the purpose of comparison, the sampling rate, the window size and the frame rate are kept identical in the computations of both LPCC and MFCC parameters. A total of 16 LPCC coefficients and 12 MFCC coefficients are extracted for each speech frame. The details of the computations of LPCC and MFCC can be found in (Atal, 1974) and (Davis and Mermelstein, 1980), respectively.

5.5. Vector quantization

After the feature extraction stage, a frame of speech samples is represented by a vector. The vector consists of 12 elements in the cases of LFPCs and MFCCs, and 16 elements in the case of LPCCs. To further compress the data for presentation to the final stage of the system, vector quantization is performed (Equitz, 1989). All the coefficients are normalized before vector quantization. A codebook of size 64 is constructed using a large set of vectors representing the features of all training samples. The division into 64 clusters is carried out according to the LBG algorithm (Furui, 1989), which is an expansion of the Lloyd's algorithm. All vectors falling into a particular cluster are coded with the vector representing the cluster.

The quality of the codebook (vector quantizer) can be measured by Distortion parameter, which is the average distance of a vector observation in the training data from its corresponding centroid of the codebook. Distortion can be reduced by increasing the codebook size. For speech recognition using MFCC, it is found in (Deller et al., 1993)

that the benefit per centroid diminishes significantly beyond the size of 32 or 64. A larger codebook also means increased computational load. From our experiments, it is also found that the performance of the proposed system does not improve significantly by extending the codebook size beyond 64.

A vector of 12 short time Log Frequency Power Coefficients for each speech frame, is assigned to a cluster by vector quantization. The vector f_n is assigned the codeword c_n^* according to the best match codebook cluster z_c using (7).

$$c_n^* = \arg\min_{1 \leqslant c \leqslant C} d(f_n, z_c) \tag{7}$$

For a speech utterance with N frames, a feature vector Y is then obtained where

$$Y = [\, c_1^* \quad c_2^* \quad \ldots \quad c_n^* \,] \tag{8}$$

5.6. Classification by hidden Markov model

Hidden Markov models (HMMs) are popular for speech recognition (Lee and Hon, 1989) and hence they are adopted for the classification of emotion in speech. According to Deller et al. (1993), the states in the HMM frequently represent identifiable acoustic phonemes in speech recognition. The number of states is often chosen to roughly correspond to the expected number of phonemes in the utterances. However, the optimal number of states is best determined through experiments as the relationship of the number of states to the performance of the HMM is very imprecise.

The structure of the HMM generally adopted for speech recognition is a left-to-right structure, since phonemes in speech follow strictly the left to right sequence. However, emotional cues contained in an utterance cannot be assumed as specific sequential events in the signal. For example, if pause is associated with the Sad emotion, there is no fixed time in the utterance for the pause to occur: it can be an event at the beginning, at the middle or at the end of the utterance. As long as pause occurs, Sadness may be considered (Yamada et al., 1995). For this reason, an ergodic model HMM is more suitable for emotion recognition since for this model, every state can be reached in a single step from every other state.

Our experimental studies show that a 4-state discrete ergodic HMM gives the best performance compared with the left-right structure. The state transition probabilities and the output symbol probabilities are uniformly initialized. A separate HMM is obtained for each emotion group during the training phase. The output symbol probabilities are smoothed with the uniform distribution to avoid the presence of too small probabilities or zero probabilities. 60% of the emotion utterances of each speaker were used to train each emotion model. After training, six HMM models are established for each speaker, one for each emotion class. Recognition tests are conducted on the remaining 40% of the utterances using the forward algorithm. The proposed system is text independent but speaker dependent as different sentences are used for each speaker and the models are trained for the individual speaker. When a test utterance is presented to the system, the utterance is scored using the forward algorithm across observed emotion models.

The probability of observation O given the Model λ, $P(O|\lambda)$, for each emotion model is calculated using (9).

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{9}$$

where $\alpha_T(i)$ is the terminal forward variable determined by the forward algorithm, and N is the number of HMM states. The model with the highest score determines the classified emotion. Details of the training and re-estimation algorithms is given in (Rabiner and Juang, 1993).

6. Experiments and analysis of results

Experiments were conducted to evaluate the performance of the proposed system. The variation of LFPC with time for utterances associated with different emotions are presented in Appendix A. In addition to the proposed system, experiments using LPCC, MFCC as feature vectors were also conducted for the purpose of comparison.

First, the performance of the system in classifying all the six basic emotions individually was assessed. The recognition rates of classification using utterances reserved for testing for all feature sets are shown in Tables 7–9. It is observed that the best average results are obtained when α is chosen as 1.4 and the center frequency of the 12th filter is chosen as 6154 Hz. The best average rates of classification for the Burmese and the Mandarin utterances are 78.5% and 75.7% respectively. Both are obtained when α is 1.4. It is noted that the classification results are relatively higher than the human classification performance of 65.8%. As for selection of feature parameters, the average percentages of classification accuracy using LPCC, MFCC and LFPC are 56.1%, 59.0%

Table 7
Classification accuracy for Burmese utterances for different values of α

Speaker	α				
	1.0	1.1	1.2	1.3	1.4
S1 (M)	70.8	70.8	79.2	75	75
S2 (M)	83.3	72.2	80.5	86.1	83.3
S3 (M)	62.5	58.3	70.8	62.5	75
S4 (F)	70.8	87.5	54.2	66.7	70.8
S5 (F)	62.5	79.2	79.2	70.8	79.2
S6 (F)	66.7	66.7	75	75	95.8
Average performance	69.4	72.5	73.2	72.7	79.9

Table 8
Classification accuracy for Mandarin utterances for different values of α

Speaker	α				
	1.0	1.1	1.2	1.3	1.4
S1 (M)	79.2	50	79.2	83.3	87.5
S2 (M)	50	58.3	62.5	70.8	70.8
S3 (M)	95.8	79.2	75	62.5	66.7
S4 (F)	66.7	70.8	79.2	79.2	75
S5 (F)	58.3	70.8	58.3	58.3	79.2
S6 (F)	75	66.7	83.3	79.2	79.2
Average performance	70.8	66	72.9	72.2	76.4

Table 9
Comparison of classification accuracy

Speaker	LPCC		MFCC		Proposed method, $\alpha = 1.4$	
	Burmese	Mandarin	Burmese	Mandarin	Burmese	Mandarin
S1 (M)	54.2	79.2	58.3	66.7	75	87.5
S2 (M)	69.4	41.7	72.2	37.5	83.3	70.8
S3 (M)	52.8	54.2	63.9	58.3	75	66.7
S4 (F)	58.3	41.7	54.2	66.7	70.8	75
S5 (F)	62.5	50	66.7	58.3	79.2	79.2
S6 (F)	58.3	50	62.5	41.7	95.8	79.2
Average performance	59.3	52.8	63	54.9	79.9	76.4
Average performance	56.1		59		78.1	

and 77.1% respectively. This shows that LFPC is a good selection as feature parameter for emotion classification of speech. To further confirm the superiority of LFPC over LPCC as feature parameters, the distributions of the coefficients for the two extreme emotions of Anger and Sadness are obtained. It is found that on the average, there is more overlap of the distributions for LPCC than for LFPC. Refer to Appendix B for further details.

The utterances of certain emotions, such as Anger, Surprise and Joy have similar acoustic features (Schubiger, 1958; O'Connor and Arnold, 1973; Williams and Stevens, 1981; Cowan, 1936; Fairbanks and Pronovost, 1939; Lynch, 1934; Frick, 1985; Murray and Arnott, 1993; Crystal, 1969; Crystal, 1975; Fonagy, 1978a,b; Fonagy and Magdics, 1963; Davitz, 1964; Williams and Stevens, 1969; Van Bezooijen, 1984; Kotlyar and Mozorov, 1976; Muller, 1960; Oster and Risberg, 1986; McGilloway et al., 1995; Trojan, 1952; Havrdova and Moravek, 1979; Huttar, 1968; Coleman and Williams, 1979; Kaiser, 1962; Scherer, 1986b; Utsuki and Okamura, 1976; Sulc, 1977; Johnson et al., 1986; Cahn, 1990). If these emotions are grouped together and treated as a group, the accuracy of classification should improve. For ease of description, this is referred to as reduced set classification.

For reduced set classification, the grouping shown in Table 10 is used. Since Anger, Surprise and Joy emotion styles are manifested similarly in energy distribution, these three emotions are put into one group (G1). Fear, Sadness and Disgust emotions also have similar spectral distributions and are put into another group (G2).

Table 10
Grouping of emotions

Emotion Group	Group members
G1	Anger, Surprise, Joy
G2	Fear, Disgust, Sadness

Two additional experiments were carried out to distinguish the groups of emotions. Experiment 1 was conducted to distinguish between emotion groups G1 and G2. Experiment 2 was carried out to distinguish between the two extreme emotions of Anger and Sadness. Results of these experiments are summarized in Table 11.

It is observed that grouping of emotions gives rise to much improved accuracy of classification. When the emotions were put into two groups, G1 and G2, the classification accuracy increases to 90% compared with the classification accuracy for six individual emotions. As expected, the classification score between Anger and Sadness is the highest (100%) since the "agitation" level of emotion in Anger differs considerably from Sadness.

As for classification performance across utterances of different languages, results presented in Tables 7–9 show that similar overall accuracy is obtained across the two languages tested, namely Burmese and Mandarin. These data suggest an existence of similar inference rules from vocal expression across languages (De Silva et al., 1998; Scherer et al., 2001).

To assess the effect of number of states for the HMM model, experiments were carried out using

Table 11
Classification accuracy (%)

Speaker	Experiment 1: G1 and G2		Experiment 2: Anger and Sadness		Experiment 3: six emotions	
	Burmese	Mandarin	Burmese	Mandarin	Burmese	Mandarin
S1 (M)	100	79.2	100	100	75	87.5
S2 (M)	86.1	79.2	100	100	83.3	70.8
S3 (M)	91.7	87.5	100	100	75	66.7
S4 (F)	87.5	91.7	100	100	70.8	75
S5 (F)	100	95.8	100	100	79.2	79.2
S6 (F)	70.8	95.8	100	100	95.8	79.2
Average performance	89.4	88.2	100	100	79.9	76.4
Average performance	88.8		100		78.1	

HMM models with one to eight states. The results are presented in Fig. 3. It is observed that HMM with 4 states delivers the best optimal performance. To further confirm that 4 is the optimal number of states required, the transition of states for different emotion utterances using 4, 5 and 6 states HMM is investigated. The details of the investigation are described in Appendix C.

The effects of frame size and size of overlapping frame were also investigated. The Table 12 shows the results of the performance of three different overlapping combinations. The proposed choice of parameters with frame size of 16 ms and overlapping duration of 9 ms gives the best performance.

The performance of the proposed system is compared with the results obtained by other approaches. The results are summarized in Table 13. These other approaches include statistical pattern recognition by Dellaert et al. (1996) and

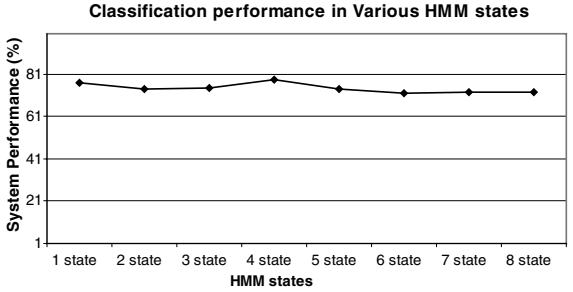

Fig. 3. Average classification performance for various number of HMM states.

Table 12
Classification accuracy for different frame sizes ($\alpha = 1.4$)

Speaker	WS = 20 ms, OD = 13 ms		WS = 24 ms, OD = 16 ms		Proposed parameters (WS = 16 ms, OD = 9 ms)	
	Burmese	Mandarin	Burmese	Mandarin	Burmese	Mandarin
S1 (M)	75	70.8	62.5	66.7	75	87.5
S2 (M)	80.5	58.3	86.1	66.7	83.3	70.8
S3 (M)	62.5	58.3	70.8	70.8	75	66.7
S4 (F)	79.2	79.2	79.2	70.8	70.8	75
S5 (F)	75	87.5	70.8	79.2	79.2	79.2
S6 (F)	79.2	79.2	83.3	75	95.8	79.2
Average performance	75.2	72.2	75.5	71.5	79.9	76.4

WS is the window size and OD is the overlapping duration.

Table 13
Comparison with other methods

Approach	Emotions classified	Average accuracy (%)
ASSESS method (McGilloway et al., 2000)[a]	Afraid, Happiness, Neutral, Sadness, Anger	55
Pattern recognition by Dellaert et al. (1996)[b]	Happiness, Sadness, Anger and Fear	79.5
Neural network by Nicholson et al. (1999)[c]	Joy, Teasing, Fear, Sadness, Disgust, Anger, Surprise, Neutral	50
Proposed method[d]	Anger, Dislike, Fear, Happiness, Sadness, Surprise	78.1

[a] Text dependent system. A total of 40 speakers (20 males and 20 females) uttered 197 passages for five emotions categories. The same set of sentences was used for all five emotions.

[b] Text dependent system. A total of five speakers uttered 1000 emotion samples. The same set of sentences was used for all four emotions. System performance was tested on the same set of sentences.

[c] Text dependent system. Each speaker uttered the same list 100 Japanese phonetically balanced words for all eight emotion types. A total of 100 speakers (50 males, 50 females) uttered 80,000 emotion utterances. The system performance was tested on the same list 100 phonetically balanced words.

[d] Text independent system. Sentences or words in different length and different language can be tested in our system. The content of emotion sentences was not fixed. A total of 12 speakers contribute 720 emotion utterances. The system performance was tested on unseen text.

neural network classification by Nicholson et al. (1999). Note that the systems differ in speaker dependency, text dependency, the number and type of emotions classified and the size of the database used. Nevertheless it provides a crude comparison of the different approaches.

7. Conclusions

In this paper, a system for classification of emotional state of utterances is proposed. The system makes use of short time LFPC for feature representation and a 4-state ergodic HMM as the recognizer.

Short time LFPC represents the energy distribution of the signal in different Log frequency bands. Spectral analysis shows that distribution of energy is dependent on emotion type and this serves as a good indication of emotion type. The coefficients also provide important information on the fundamental frequency of speech. By integrating the coefficients with an HMM recognizer, the variation of F0 values and speaking rate are taken into account. The proposed system is able to take into consideration the most important parameters in speech emotion recognition.

The results of the experiments show that average accuracy of 77.1% and best accuracy of 89% can be achieved in classifying the six basic emotions individually. The results are better than accuracy of 65.8% achieved by human assessment. Higher accuracy can be achieved if the six basic emotions of similar nature are merged into fewer groups. This indicates that the proposed system with short time LFPC as the indicators of emotional content and HMM as the classifier do serve as a viable approach for the classification of emotions of speech.

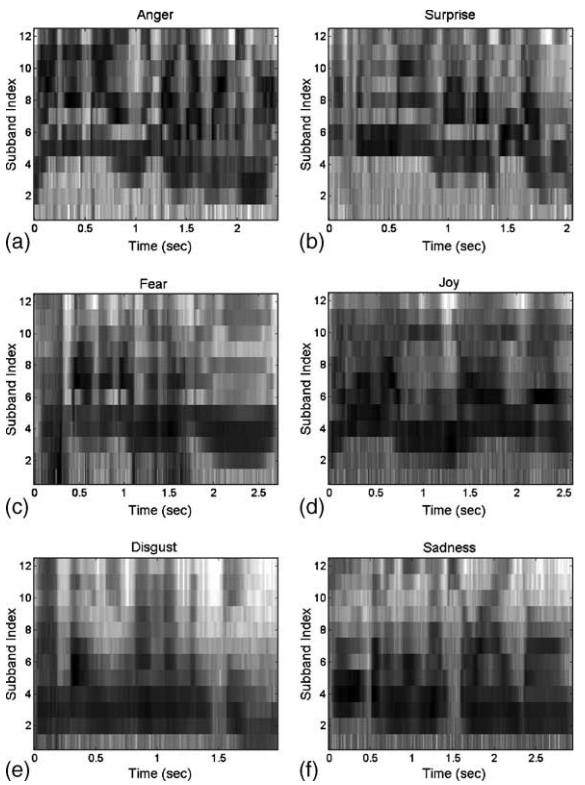

Fig. 4. Log energy spectrum of utterances in Burmese (female speaker).

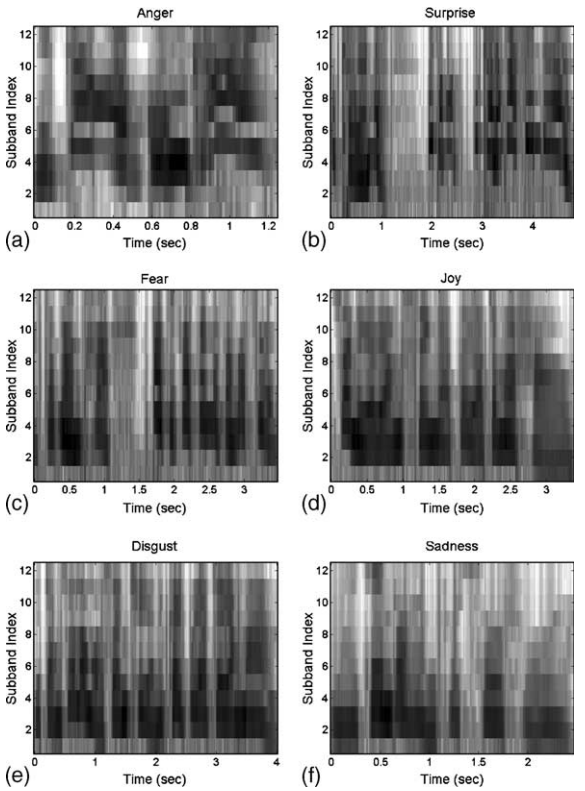

Fig. 5. Log energy spectrum of utterances in Mandarin (female speaker).

Appendix A. Variation of LFPC with time

The variations of LFPC with time for utterances associated with different emotions of Burmese and Mandarin speakers are displayed in Figs. 4–7. α value in these cases is set at 1.4. The ordinate gives the sub-band index (which represents frequency) and time is represented on the abscissa. Higher energy is indicated by the darker print levels. From the figures, it can be observed that the pattern of distribution of spectral energy is different for utterances associated with different emotions. For utterances associated with Anger and Surprise emotions, the average energy contents are higher than those for utterances associated with Sadness and Fear. It can also be observed that in general, for Anger and Surprise, the energy is comparatively higher in the higher bands. On the other hand, for Disgust and Sadness, the energy concentrates at the lower bands.

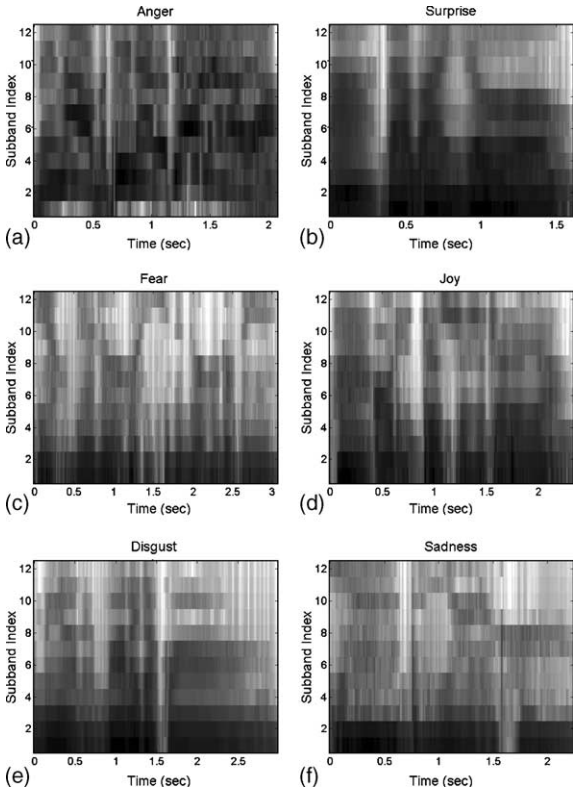

Fig. 6. Log energy spectrum of utterances in Burmese (male speaker).

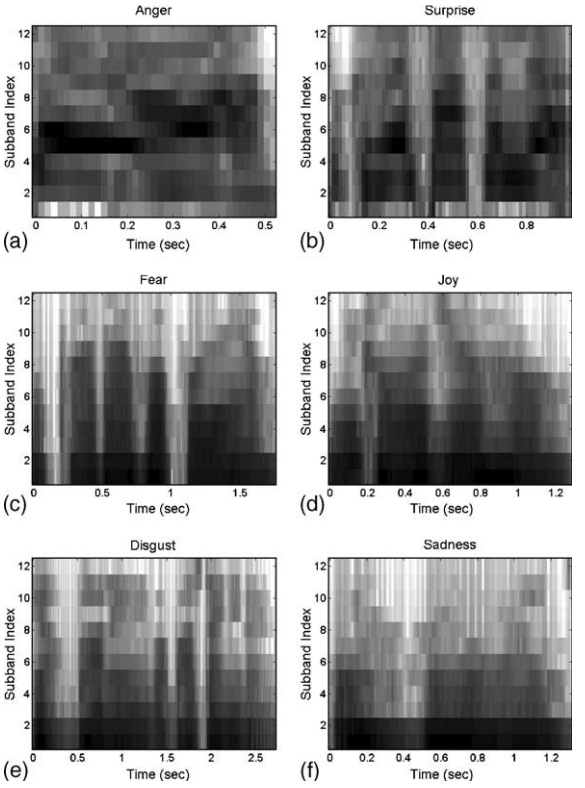

Fig. 7. Log energy spectrum of utterances in Mandarin (male speaker).

For Fear and Joy, the energy envelope is sandwiched between the other two. The rate of changes of the spectral energy is also faster for Anger and Surprise and slower for Dislike and Sadness.

The average intensities in the 12 sub-bands over time for the two extreme emotions of Anger and Sadness are also computed and shown in Fig. 8 for four different speakers. It can be observed that Anger (high arousal emotion) has higher intensity values in higher frequency bands while Sadness (low arousal emotion) has higher intensity values in lower frequency bands.

Appendix B. Distributions of feature parameters

To compare the performance of LFPC and LPCC as feature parameters, the distributions of the coefficients for utterances of the two extreme

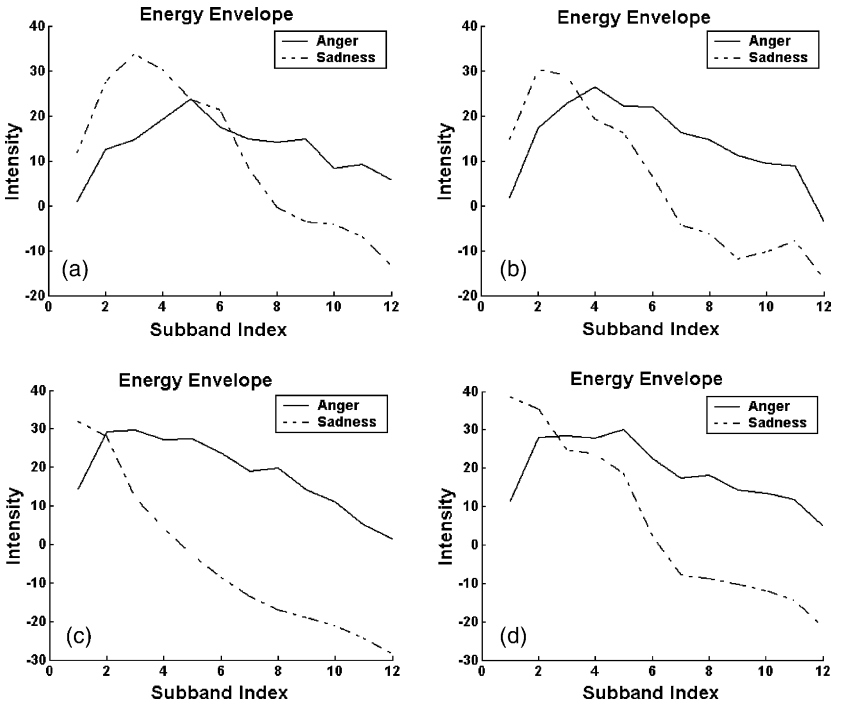

Fig. 8. Comparison of intensity values for Anger and Sadness emotions: (a) Burmese female, (b) Mandarin female, (c) Burmese male and (d) Mandarin male speakers.

emotions of Anger and Sadness are obtained. The coefficients are determined as stated in the main text with α set at 1.4. For LFPC, there are 12 essential coefficients. The silence intervals of the utterances are first removed. The normalized histograms of the coefficient values using utterances from Burmese male speakers are depicted in Fig. 9. The total length of the utterances used is approximately 8 seconds. To measure the degree of overlap between the two distributions, the Elias coefficient (Elias, 1975) is computed. Mathematically, the Elias coefficient, M, is calculated as follows.

$$M = \int_{-\infty}^{+\infty} |p_1(x) - p_2(x)| \, \mathrm{d}x \tag{B.1}$$

where $p_1(x)$ and $p_2(x)$ are the probability densities associated with the two distributions. An Elias coefficient of 2 indicates a complete separation of the two distributions and a value of 0 means complete overlap.

The Elias coefficients for the 12 coefficients are calculated and shown in Fig. 10. The average Elias coefficient for all 12 LFPC is also computed.

For LPCC, there are 12 essential coefficients. As for the LFPC, the normalized histograms of each of the 12 coefficients for the two extreme emotions are obtained and the Elias coefficients are computed. These are given in Figs. 11 and 12.

Comparing the mean Elias coefficients for LFPC and LPCC, it can be concluded that on the average, the degree of separation of the distributions is greater for LFPC than for LPCC.

Appendix C. Number of states for HMM

To assess the effect of the number of states for the HMM model, experiments are carried out using HMM models with different number of states. The state transition diagrams for the utterances of the Disgust and Sadness emotions using 4-, 5- and 6-state HMMs are given in Figs. 13 and 14 respectively. From these figures, it can be observed

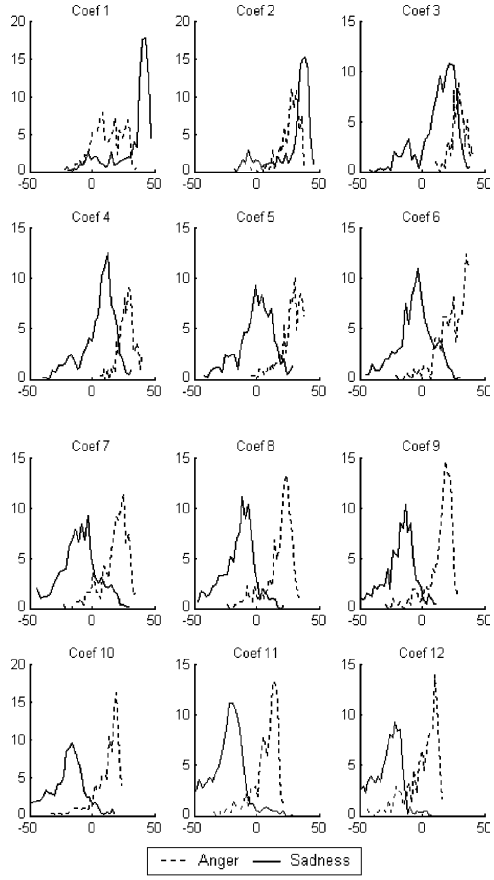

Fig. 9. Distribution of 12 LFP coefficient values (Burmese male speakers). The abscissa represents 'Log frequency power coefficient values' and the ordinate represents 'Percentage of Coefficients'.

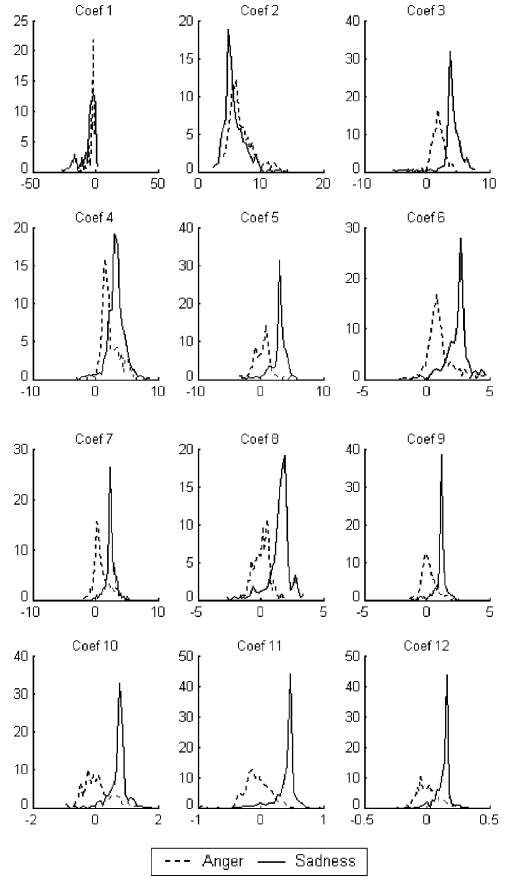

Fig. 11. Distribution of 12 LPC coefficient values (Burmese male speakers). The abscissa represents 'Coefficient Values' and the ordinate represents 'Percentage of Coefficients'.

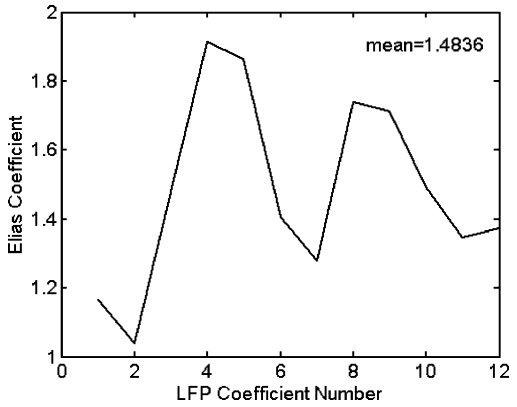

Fig. 10. Elias coefficients between 'Anger' and 'Sadness' emotions (Burmese male speakers) using LFP coefficients.

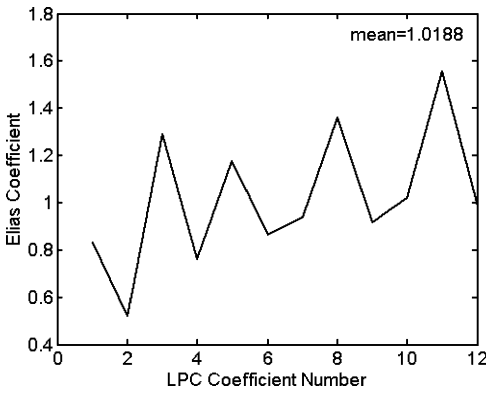

Fig. 12. Elias coefficients between 'Anger' and 'Sadness' emotions (Burmese male speakers) using LPC coefficients.

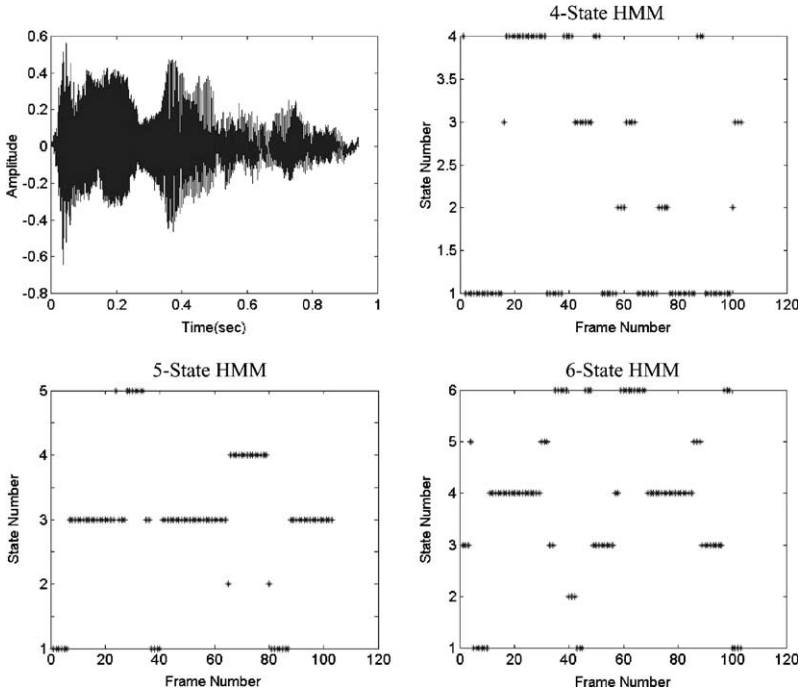

Fig. 13. Waveform and state transition diagrams of 'Disgust' utterance.

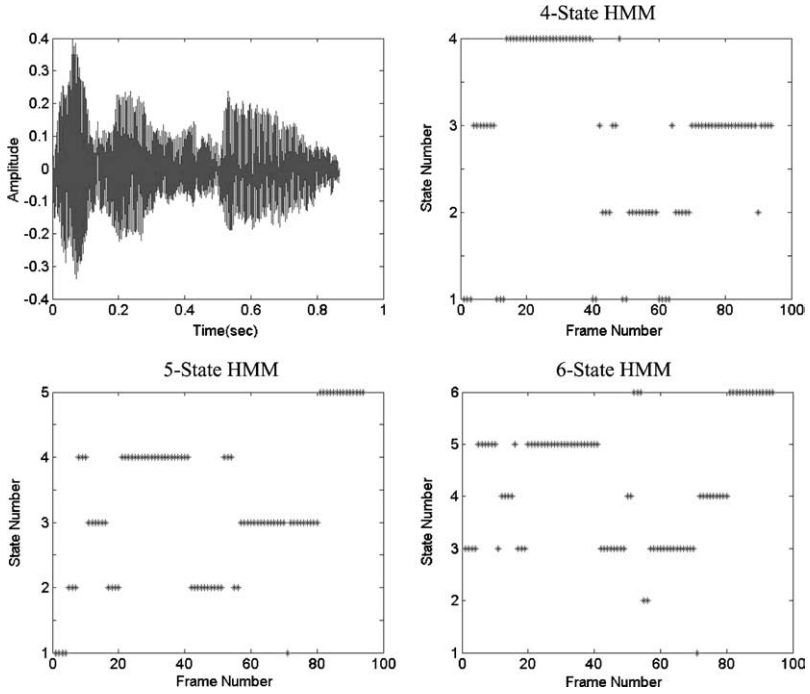

Fig. 14. Waveform and state transition diagrams of 'Sadness' utterance.

that a high percentage of the feature vectors stay in four of the states if there are more than 4 states. We may regard the states as representing the spectral energy contents in this instance.

References

Arnold, M.B., 1960. Emotion and Personality. Physiological Aspects, Vol. 2. Columbia University Press, New York.

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Amer. 55 (6), 1304–1312.

Becchetti, C., Ricotti, L.P., 1998. Speech Recognition Theory and C++ Implementation. John Wiley & Sons, New York.

Cahn, J.E., 1990. The generation of affect in synthesized speech. J. Amer. Voice I/O Soc. 8, 1–19.

Cairns, D.A., Hansen, J.H.L., 1994. Nonlinear analysis and classification of speech under stressed conditions. J. Acoust. Soc. Amer. 96 (6), 3392–3400.

Coleman, R., Williams, R., 1979. Identification of emotional states using perceptual and acoustic analyses. In: Lawrence, V., Weinberg, B. (Eds.), Care of the Professional Voice, Vol. 1. The Voice Foundation, New York.

Cornelius, R., 1996. The Science of Emotion. Prentice-Hall, Englewood Cliffs, NJ.

Cowan, M., 1936. Pitch and Intensity Characteristics of Stage of Speech. Arch. Speech, suppl. to Dec. issue.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human–computer interaction. IEEE Sig. Proc. Mag. 18 (1), 32–80.

Crystal, D., 1969. Prosodic Systems and Intonation in English. Cambridge University Press, London, UK.

Crystal, D., 1975. The English Tone of Voice. Edward Arnold, London, UK.

Darwin, C., 1965. The Expression of Emotions in Man and Animals. John Murray, Ed., 1872. Reprinted by University Press.

Davis, M., College, H., 1975. Recognition of Facial Expressions. Arno Press, New York.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357–366.

Davitz, J.R. (Ed.), 1964. The Communication of Emotional Meaning. McGraw-Hill, New York.

De Silva, L.C., Miyasato, T., Nakatsu, R., 1998. Use of multimodal information in facial emotion recognition. IEICE Trans. Inf. Syst. E81-D (1), 105–114.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing Emotion in Speech. Fourth International Conference on Spoken Language Processing 3, 1970–1973.

Deller, J.R., Proakis, J.G., Hansen, J.H.L., 1993. Discrete-Time Processing of Speech Signals. Macmillan Pub. Co., Toronto.

Ekman, P., 1973. Darwin and Facial Expressions. Academic, New York.

Ekman, P., Friesen, W., 1975. Unmasking the Face. Prentice-Hall, Englewood Cliffs, NJ.

Elias, N.J., 1975. New Statistical Methods for Assigning Device Tolerances. Proc. IEEE Int. SYmp. Ccts. Sys., 1975, Newton, MA, USA, pp. 329–332.

Equitz, W.H., 1989. A new vector quantization clustering algorithm. IEEE Trans. Acoust. Speech Signal Process. 37 (10), 1568–1575.

Fairbanks, G., Pronovost, W., 1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monograph 6, 87–104.

Fonagy, I., 1978a. A new method of investigating the perception of prosodic features. Language and Speech 21, 34–49.

Fonagy, I., 1978b. In: Sundberg, J. (Ed.), Emotions, Voice and Music in Language and Speech, Vol. 21, pp. 34–49.

Fonagy, I., Magdics, K., 1963. Emotional patterns in intonation and music. Z. Phonet. Sprachwiss. Kommunikationsforsch 16, 293–326.

Fox, N.A., 1992. If it's not left it's right. Amer. Psychol. 46, 863–872.

Frick, R., 1985. Communicating Emotion: The Role of Prosodic Features. Psychol. Bull. 97 (3), 412–429.

Furui, S., 1989. Digital Speech Processing, Synthesis and Recognition. Marcel Dekker, New York.

Havrdova, Z., Moravek, M., 1979. Changes of the voice expression during suggestively influenced states of experiencing. Activitas Nervosa Superior 21, 33–35.

Huttar, G.L., 1968. Relations between prosodic variables and emotions in normal american english utterances. J. Speech Hearing Res. 11, 481–487.

Johnson, W., Emde, R., Scherer, R., Klinnert, M., 1986. Recognition of emotion from vocal cues. Arch. Gen. Psych. 43, 280–283.

Kaiser, L., 1962. Communication of affects by single vowels. Synthese 14, 300–319.

Kotlyar, G., Mozorov, V., 1976. Acoustic correlates of the emotional content of vocalized speech. J. Acoust. Acad. Sci. USSR 22, 208–211.

Lazarus, R.S., 1991. Emotion Adaptation. Oxford University Press, New York.

Lee, K.F., Hon, H.W., 1989. Speaker-independent phone recognition using Hidden Markov Models. IEEE Trans. Acoust. Speech Signal Process. 37 (11), 1641–1648.

Lynch, G.E., 1934. A phonophotographic study of trained and untrained voices reading factual and dramatic material. Arch. Speech 1, 9–25.

McGilloway, S., Cowie, R., Douglas-Cowie, E., 1995. Prosodic signs of emotion in speech: preliminary results from a new technique for automatic statistical analysis. In: Proc. XIIIth Int. Congr. Phonetic Sciences, Vol. 1. Stockholm, Sweden, pp. 250–253.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S., 2000. Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark. ISCA Workshop on Speech and Emotion, Belfast.

Muller, A., 1960. Experimentelle Untersuchungen zur stimmlichen Darstellung von Gefuehlen [Experimental Studies on Vocal Portrayal of Emotion], Ph.D. dissertation, Univ. Gottingen, Germany.

Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoust. Soc. Amer. 93 (2), 1097–1108.

Nicholson, J., Takahashi, K., Nakatsu, R., 1999. Emotion recognition in speech using neural networks. 6th International Conference on Neural Information Processing, ICONIP '99, Vol. 2, pp. 495–501.

Oatley, K., Johnson-Laird, P., 1995. Communicative theory of emotions: Empirical test, mental models and implications for social interaction. In: Martin, L., Tessler, A. (Eds.), Goals and Affect. Erlbaum, Hillsdale, NJ.

O'Connor, J.D., Arnold, G.F., 1973. Intonation of Colloquial English, second ed. Longman, London, UK.

Oster, A., Risberg, A., 1986. The Identification of the Mood of A Speaker by Hearing Impaired Listeners. Speech Transmission Lab. Quarterly Progress Status Report 4, Stockholm, pp. 79–90.

Otaley, K., Jenkis, J.M., 1996. Understanding Emotions. Blackwell, Oxford, UK.

Plutchik, R., 1994. The Psychology and Biology of Emotion. Harper Collins, New York. p. 58.

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

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice Hall, Englewood Cliffs, NJ.

Scherer, K.R., 1984. On the nature and function of emotion: A component process approach. In: Scherer, K.R., Ekman, P. (Eds.), Approaches to Emotion. Erlbaum, Hillsdale, NJ.

Scherer, K.R., 1986a. Vocal effect expression: A review and a model for future research. Psychol. Bull. 99, 143–165.

Scherer, K.R., 1986b. Vocal affect expression: A review and a model for future research. Psychol. Bull. 99, 143–165.

Scherer, K.R., Banse, R., Wallbott, H.G., 2001. Emotion inferences from vocal expression correlate across languages and cultures. J. Cross-Cultural Psychol. 32 (1), 76–92.

Scherer, K., Ekman, P., 1984. Approaches to Emotion. Lawrence Erlbaum Associates, Mahwah, NJ.

Schubiger, M., 1958. English Intonation. Its Form and Function. Niemeyer, Tubingen, Germany.

Sulc, J., 1977. Emotional changes in human voice. Activitas Nervosa Superior 19, 215–216.

Trojan, F., 1952. Der Ausdruck der Sprechstimme. W. Maudrich, Wien-Dusseldorf, Germany.

Utsuki, N., Okamura, N., 1976. Relationship Between Emotional State and Fundamental Frequency of Speech. Rep. Aeromedical Laboratory. Japan Air Self-Defense Force 16, 179–188.

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

Williams, C.E., Stevens, K.N., 1969. On determining the emotional state of pilots during flight: An exploratory study. Aerospace Med. 40, 1369–1372.

Williams, C.E., Stevens, K.N., 1981. Vocal correlates of emotional states. In: Darby, J.K. (Ed.), Speech Evaluation in Psychiatry. Grune and Stratton, Inc., pp. 189–220.

Yamada, T., Hashimoto, H., Tosa, N., 1995. Pattern recognition of emotion with neural network. In: Proc. 1995 IEEE IECON 21st Internat. Conf. Industrial Electronics, Control, and Instrumentation, Vol. 1, pp. 183–187.