EMPIRICAL MODE DECOMPOSITION BASED WEIGHTED FREQUENCY FEATURE FOR SPEECH-BASED EMOTION CLASSIFICATION

Vidhyasaharan Sethu[1,2]*, Eliathamby Ambikairajah*[1,2] *and Julien Epps*[1]

[1]The School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052, Australia
[2]National Information Communication Technology (NICTA),
Australian Technology Park, Eveleigh 1430, Australia

ABSTRACT

This paper focuses on speech based emotion classification utilizing acoustic data. The most commonly used acoustic features are pitch and energy, along with prosodic information like rate of speech. We propose the use of a novel feature based on instantaneous frequency obtained from the speech, in addition to the aforementioned features, in order to take into account the vocal tract parameters as well as vocal chord excitation. The proposed features employ the recently emerged empirical mode decomposition to decompose speech into AM-FM signals that are symmetric about zero and suitable for Hilbert transformation to extract the instantaneous frequency. The proposed features provide a relative increase in classification accuracy of approximately 9% when appended to established acoustic features.

Index Terms— Emotion classification, instantaneous frequency, empirical mode decomposition, hidden Markov models, front-end processing.

1. INTRODUCTION

Automatic emotion classification has gained increasing attention from researchers over the past few years due to its potentially broad range of applications, including computer-based tutoring systems, tele-monitoring of patients and call centre services that can automatically transfer angry customers to human operators [1].

Our focus is the development of a system that can detect the emotional state of a person based on speech. The system considered in this paper does not make use of semantic or linguistic information and as such does not make use of language models. Such systems rely solely on prosodic and/or spectral features such as pitch, intensity, speech rate, cepstral coefficients, group delay [1-5].

Among features proposed for emotion recognition, those derived from pitch and energy are the most popular for a speaker-independent emotion classification system, where data from target speakers are not available for training. These features characterize the state of the vocal chords, but do not provide any information as to the state of the vocal tract. On the other hand, features based on Mel frequency cepstral coefficients (MFCCs) and group delay of the all pole filter modelling the vocal tract, while useful in a speaker-dependent emotion detection system, are outperformed by vocal chord parameters in a speaker-independent system [5]. This is most likely due to the non-trivial differences in the vocal tract characterisations for different speakers. Thus, a feature vector that is derived from the speech spectrum, but excludes details that vary between different speakers will be useful for a speaker-independent emotion classifier. One way of condensing the information contained in the speech spectrum is to obtain broad measures of the spectral magnitude distribution from the DFT, such as the spectral slope [4] or spectral centroid [6]. However, the time resolution of such features will be constrained by the DFT window length and these are crude approximations during non-stationary segments of speech.

This problem can be overcome using an estimate of the instantaneous frequency. The recently pioneered empirical mode decomposition (EMD) [7] can be used to represent the speech signal as a sum of zero-mean AM-FM components which then allow for the definition of a positive instantaneous frequency for each component based on the Hilbert transform. We propose the use of a weighted frequency feature based on these component instantaneous frequencies for a speaker-independent emotion classification system.

2. EMD BASED INSTANTANEOUS FREQUENCY

Any real-valued signal can be written as an analytic signal by setting it as the real part of the analytic signal and its Hilbert transform as the imaginary part of the analytic signal

$$z(t) = x(t) + iy(t) \, , \tag{1}$$

where $x(t)$ is the real valued signal and $y(t)$ is the Hilbert transform of $x(t)$.

From the analytic signal, the instantaneous phase can be obtained and the time derivative of the instantaneous phase is then defined as the instantaneous frequency. The complex

1-4244-1484-9/08/$25.00 ©2008 IEEE

analytic function also allows for the definition of instantaneous amplitude.

$$\phi(t) = \tan^{-1}\left(\frac{y(t)}{x(t)}\right) \tag{2}$$

$$\theta(t) = \frac{d\phi(t)}{dt} \tag{3}$$

$$a(t) = \sqrt{x^2(t) + y^2(t)} \tag{4}$$

where $\phi(t)$ is the instantaneous phase, $\theta(t)$ is the instantaneous frequency and $a(t)$ is the instantaneous amplitude.

A problem for most methods of instantaneous frequency estimation occurs when sudden changes in the amplitude or frequency of the signal result in the instantaneous frequency paradoxically taking negative values [8]. The necessary conditions for a meaningful definition of instantaneous frequency based on the analytic representation of the signal are that the signal is symmetric with respect to the local zero mean, and has the same number of extrema and zero crossings [7]. Functions satisfying these conditions are referred to as intrinsic mode functions (IMF) by Huang *et al.* [7]. The empirical mode decomposition (EMD) [7] enables any signal to be written as a sum of a few intrinsic mode functions and in some cases a monotonic residue that represents the overall trend of the signal. The empirical mode decomposition process begins by extracting the first intrinsic mode function, which consists of oscillations on the smallest scale, locally by a sifting process. This IMF is then subtracted from the signal and the process is iterated until all possible intrinsic mode functions have been extracted and only a monotonic residue is left (Fig. 1.).

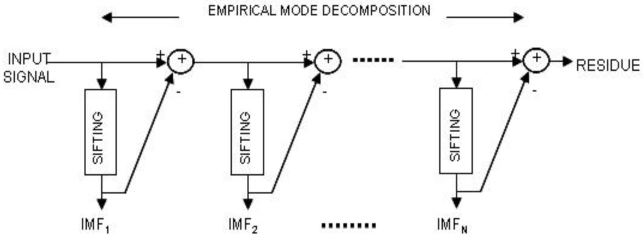

Fig. 1. Overview of the empirical mode decomposition

Due to the present lack of a mathematical framework for the EMD, there are limitations to the study of its properties. In the following section we investigate its application to speech signals using empirical methods.

3. WEIGHTED FREQUENCY FEATURE

3.1. Investigation of IMFs in Speech

As seen in equations (1) to (4), the instantaneous amplitude and frequency can be used to characterise the spectral content of a signal such as speech. Typically, speech signals sampled at 22050 Hz contain between 14 and 19 IMFs and the instantaneous amplitudes and frequencies derived from all these intrinsic mode functions (together with the residue) contain all the information present in the signal. In informal experiments it was observed that for speech signals in general approximately the first five modes (IMFs) contained most of the perceptually significant information.

An investigation of this observation, was conducted by measuring the PESQ scores of speech reconstructed from the *M* most significant modes together with mean IMF frequencies and mean IMF energies from over 9 min of 22050 Hz sampled speech. Results from this experiment, together with informal listening tests, showed that speech reconstructed from the first five modes (IMFs) was of sufficiently high quality for classification tasks. It was also observed that the spectral region addressed by IMFs beyond 5 is very small and likely to be correlated with pitch information. Thus, only the first five modes were used in all experiments reported in this paper. Average PESQ scores obtained for speech signals reconstructed with different number of IMFs, and the mean instantaneous frequency of each intrinsic mode function are shown in Fig. 2.

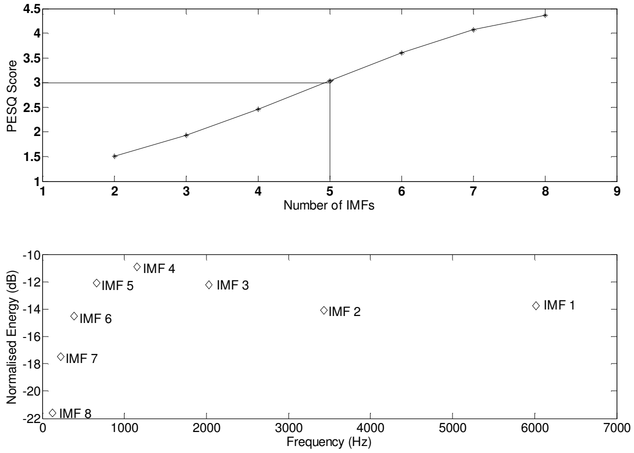

Fig. 2. (a) Average PESQ scores for reconstructed speech using different number of IMFs; (b) Mean Instantaneous Frequency and mean energy for first 8 IMFs

3.2. Emotion Characterization using EMD-Based IFs

The speech spectrum changes significantly according to the phoneme being uttered, the speaker and the emotional state of the speaker, among other factors. The changes in the instantaneous frequencies due to changes in speech content and the different vocal tract characteristics of different speakers makes using them directly as features for an emotion classifier impossible. We propose using the weighted average of the instantaneous frequencies of the first five modes, with the instantaneous amplitudes acting as the weights, as a feature. The weighted frequency, *wf*[*n*] is defined as follows:

$$wf[n] = \frac{\sum_{m=1}^{M} a_m[n]\theta_m[n]}{\sum_{m=1}^{M} a_m[n]}, \tag{5}$$

where $a_m[n]$ and $\theta_m[n]$ are the instantaneous amplitude and frequency of the m^{th} IMF and $M = 5$ in our experiments.

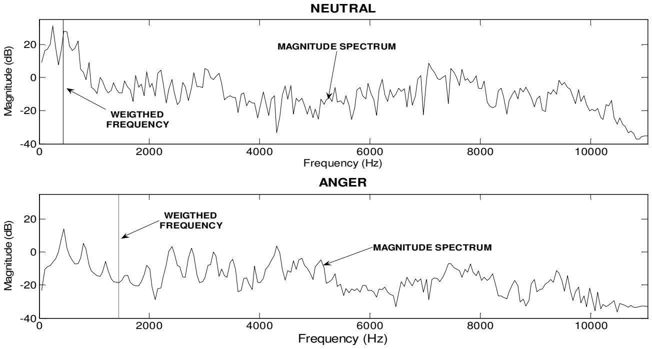

Fig. 3. Magnitude spectra and average weighted frequency values for 20ms frames of speech of phoneme /*aa*/ for two emotions: (a) Neutral; (b) Anger.

During the computation of pitch, which is one of the most commonly used features in speaker-independent emotion classifiers, information pertaining to the state of the vocal tract is discarded since pitch is a characteristic of the vocal chord vibration. Weighted frequency, *wf*[*n*] on the other hand, is computed from the speech signal without any pre-processing and takes into account the spectral shaping imposed by the vocal tract onto the vocal chord excitations. The weighted frequency is indicative of the energy distribution in the speech spectrum (Fig. 3), taking small values when most of the energy is concentrated in the low frequencies (first formant) and larger values when higher frequencies (higher formants) contain more energy. This information is useful for discriminating between emotions, as shown in the above figure, and similar differences in the weighted frequency due to emotional states were observed in other phonemes as well.

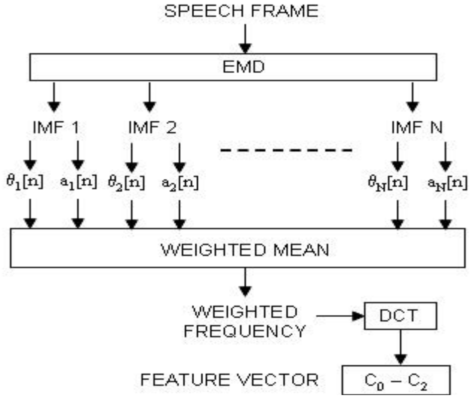

Fig. 4. Weighted frequency feature extraction.

In our experiments, a weighted frequency feature was computed from *wf*[*n*] (not directly feasible as a feature) for 40ms frames using the EMD sifting process and stopping conditions suggested by Rilling *et al.* [9]. The discrete cosine transform of this weighted frequency was then obtained and the first three coefficients were selected as a feature vector to represent that frame of data (Fig. 4.).

4. EMOTION CLASSIFICATION SYSTEM

4.1. Front-End

For our system, we employed the features (ZEPS) proposed by Huang *et al.* [4] with one change. The energy slope feature, a ratio of the energy contained in the high frequency region of the spectrum (> 1 kHz) to that of the low frequency region of the spectrum (< 1 kHz), was replaced by the weighted frequency feature (WF) proposed in this paper. The WF is representative of the spectral region containing the most energy and, we hypothesize, is more informative than the energy slope feature which only gives the relative distribution of energy between high and low frequency regions. Thus the 3-dimensional weighted frequency features were concatenated with pitch, energy and zero crossing rate (ZCR) to give a 6-dimensional feature vector "ZEP+WF" per frame.

All features were computed within 40 ms frames overlapped by 30ms. Rectangular windows were used since pitch, energy and weighted frequency estimation do not allow for the use of a tapered window and the ZCR is unaffected by window choice. Sequences of the 6-dimensional features computed for 10 consecutive frames were then passed to a speaker-independent HMM-based sequential back-end, capable of accounting for temporal variations within the sequence. Thus, classifier decisions were based on 130ms of speech. It should be noted that pitch estimates can be made only for voiced speech, thus only those sequences spanning voiced speech were used in both training and testing of the classifier. Previously, we showed that a modified feature warping technique can be used to reduce inter-speaker variability and improve the accuracy of a speaker independent emotion classifier [10], and this was applied to all features in our experiment.

4.2. Back-End

It has been suggested that sequential classifiers (such as HMM-based classifiers) are better suited for the task of classifying emotions than other commonly used non-sequential classifiers such as support vector machines and decision trees [4]. Observations from our preliminary investigations into back-end configurations tend to agree with this suggestion. In this paper we use an HMM-based classifier and a feature vector devoid of delta features, similar to the system used in [4]. Each state of the HMM is modelled by a Gaussian mixture model (GMM). For each emotion, a hidden Markov model is trained and the emotion corresponding to the model best matching the incoming test sequence is chosen as the emotion for that sequence. In contrast to the system in [4], we use shorter sequences in order to increase the number of tests, enabling us to estimate reliable statistics at the cost of slightly reduced accuracy. We do this since our aim here is to show that the proposed weighted frequency features contain information rather than to build a complete emotion classification system. In our experiments all emotions were modelled by 4-state HMMs, with each state represented by a GMM containing 4 mixtures, which provided the best trade-off between generalisation and accurate modelling of the feature distributions during empirical work.

5. EXPERIMENTS

For our experiments we used the LDC Emotional Prosody Speech corpus [11], comprising speech from professional actors trying to express emotions while reading short phrases consisting of dates and numbers. There is therefore no semantic or contextual information available. The entire database consists of 7 actors expressing 15 emotions for around 10 utterances each. When recording the database, actors were instructed to repeat a phrase as many times as necessary until they were satisfied the emotion was expressed and then move onto the next phrase. Only the last instance of each phrase was used in this experiment.

The system described in section 3 (Figure 4) was implemented with different features in order to judge the performance of the proposed features. The experiments were repeated 7 times in a ‘leave-one-out’ manner, using data from each of the 7 speakers as the test set in turn and the data from the other 6 as the training set. Experiments for a five-emotion classification problem involving Neutral, Anger, Happiness, Sadness and Boredom were performed. Classification accuracies obtained using ZEP+WF are compared with those obtained from ZEPS [4] and cepstral features in Table 1. From these results, it can be seen that weighted frequency does indeed contain more discriminative information than the energy slope. Moreover, the best performance (averaged over all emotions) is achieved when the WF features are used in combination with ZEP features.

6. CONCLUSION

This paper presents a novel feature for use in a speaker-independent multi-class speech based emotion classification system. We obtain estimates of multiple instantaneous frequencies and amplitudes present in speech from the analytic representation of the intrinsic mode functions obtained from an empirical mode decomposition of the speech signal. A weighted mean frequency is then computed and discrete cosine transformed to obtain a compact representation. The weighted frequency feature emphasizes the region of the spectrum containing the largest amount of energy at each instant, which in turn is affected by the emotional state of the speaker. Evaluation results show that the addition of the proposed weighted frequency features to the front-end of a five-emotion classifier produces an increase in the classification accuracy. Current research is focused on an attempt to better utilise the information contained in the instantaneous frequencies.

Table 1. Comparison of five class speaker-independent emotion classification accuracies (%), evaluated on the LDC Emotional Prosody database

Emotion	Energy Slope (S) alone	WF alone	ZEPS	MFCC	ZEP + WF
Neutral	43.1	**44.5**	33.0	36.1	37.6
Anger	54.4	66.4	73.3	64.4	**75.7**
Sadness	35.5	30.7	35.6	36.8	**41.2**
Happiness	21.7	21.9	**42.0**	26.0	40.0
Boredom	9.8	30.0	36.2	**39.6**	38.5
Mean	30.2	36.2	41.1	38.0	**44.7**

7. REFERENCES

[1] Yacoub, S., Simske, S., Lin, X., and Burns, J., “Recognition of Emotions in Interactive Voice Response systems”, in *Proc. EUROSPEECH*, pp. 729-732, 2003.

[2] Verceridis, D., Kotropoulus, C., and Pitas, I., “Automatic Emotional Speech Classification”, in *Proc. IEEE ICASSP*, vol. 1, pp. I- 593-596, 2004.

[3] Schuller, B., Rigoll, G., and Lang, M., “Hidden Markov Model based Speech emotion recognition”, in *Proc. IEEE ICASSP*, vol. 2, pp. II- 1-4, 2003.

[4] Huang, R., and Ma, C., “Towards a Speaker-Independent Real-time Affect Detection System”, in *Proc. 18th Int. Conf. on Pattern Recognition* (ICPR’06), vol. 1, pp. I- 1204-1207, 2006.

[5] Sethu, V., Ambikairajah, E., and Epps, J., “Group Delay Features for Emotion Detection,” in *Proc. INTERSPEECH,* pp. 2273-2276, 2007.

[6] Scheirer, E., and Slaney, M., “Construction and evaluation of a robust multifeature speech/music discriminator,” in *Proc. IEEE ICASSP,* vol. 2, pp. 1331-1334, 1997.

[7] Huang, N.E., Shen, Z., Long, S.R., Wu, M.L., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., and Liu H.H., “The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis,” in *Proc. Roy. Soc. London A,* pp. 903-995, 1998.

[8] Cohen, L., *Time-frequency analysis,* Englewood Cliffs, N.J.: Prentice-Hall, 1995.

[9] Rilling, G., Flandrin, P., and Goncalves, P., “On empirical mode decomposition and its algorithms,” in *Proc. IEEE EURASIP Workshop Nonlinear Signal Image Processing,* Italy, 2003.

[10] Sethu, V., Ambikairajah, E., and Epps, J., “Speaker normalisation for speech based emotion detection,” in *Proc. 15th Int. Conf. Digital Signal Processing,* pp. 611-614, 2007

[11] Emotional Prosody Speech corpus, Linguistic Data Consortium, University of Pennsylvania, PA, USA, http://www. ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28