IMPROVING SPEECH EMOTION RECOGNITION USING FREQUENCY AND TIME DOMAIN ACOUSTIC FEATURES

Simina Emerich Eugen Lupu

Technical University of Cluj-Napoca, Communications Dept.
Simina.Emerich@com.utcluj.ro

ABSTRACT

The recognition of the internal emotional state of a person plays an important role in several human-related fields. The present approach proposes the classification of 7 emotions (happiness, anger, fear, boredom, sadness, disgust and neutral) by using the speech signal. Different wavelet decomposition structures are used for feature vector extraction. The models were trained and tested with a Support Vectors Machine classifier.

Keywords: *emotional state, speech, SVM, Wavelet Transform*

1. INTRODUCTION

Several types of information may be provided to the listener by the speech signal. By speech, a message which is independent to the speaker is transmitted via the words. At the other levels, speech transmits other information about the speaker: health state, emotion, socioeconomic background, language employed, gender, stress and accent.

While the goal of speech recognition is recognizing words from the speech signal, the aim of speaker recognition systems is to extract the information from the speech signal depending on speaker identity.

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch average	Very much higher	Much higher	Slightly slower	Very much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice quality	Breathy, chest tone	Breathy, blaring	Resonant	Irregular voicing	Grumbled chest tone
Pitch change	Abrupt, on stressed syllables	Smooth, upward inflections	Downward inflections	Normal	Wide downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

Table 1. *Summary of most general correlates to emotion in speech*

Emotion recognition is gaining attention due to the widespread applications into various domains: detecting frustration, disappointment/tiredness, surprise/amusement etc. Applications of speech emotion recognition include psychiatric diagnosis, intelligent toys, lie detection, learning environment, educational software etc. Emotion detection may be also useful in the synthesis of emotional speech or in the finding of the correlation between prosody and emotional state.

Emotion can affect speech in different ways: consciously, unconsciously and through the autonomous nervous system. The most fundamental problem is to answer the question: what are the features which define emotion? A highly qualitative correlation between emotion and some speech features is presented in Table 1 [Murray and Arnott, 1993].

A proper choice of feature vectors is one of the most important tasks. There are many approaches towards automatic recognition of emotion in speech by using different feature vectors. There are long-time and short-time vector features. The long-time ones are estimated over the entire length of the utterance, while the short-time ones are determined over a window of usually less than 100 ms.

The long-time approach is favoured by contemporary research as emotions are identified more efficiently by these features. The wider pitch contour which is usually present in interrogative sentences and which implies a larger pitch standard deviation of the interrogative phrase consists of an argument for the use of short-time features. Also, prosodic features are to always be included into the vector, these including pitch, intensity and durations [1].

A very interesting study on the features employed in emotion detection from speech is presented in [1]. The employed features may be acoustic or linguistic. The number of acoustic features can vary from hundreds to thousands and the number of linguistic features may vary from tens to hundreds [1].

The feature type used in different approaches may be acoustic: duration, energy, pitch, spectrum, cepstrum (MFCC features), voice quality, wavelets or linguistic: bag of words (BOW), part of speech (POS), higher semantics (SEM) and varia (disuencies/non-verbals such as breathing or laughter) [2].

The following types of functionals have been mentioned in literature: sequential and combinatorial: (e. g. ratio of mean of two different Low Level Descriptors), extremes (min/max by value, min/max position, range, and slope min/max, as well as on/off-position), means: first moment by arithmetic mean and centroid, percentiles (quartiles 1/2/3, quartile ranges lower/upper/total and other percentiles), higher statistical moments (standard deviance, variance, skewness, kurtosis, length, and zero-crossing-rate), specific

functions (distributional, spectral, regressional) or non attributable (features cannot be attributed un-equivocally to one of the other types) [2].

Emotion recognition results are far more difficult to quantify than those provided by speech recognition systems. The results provided by the emotion recognition systems are comparable to human emotion recognition accuracy 70-75%.

An emotional speech data collection is undoubtedly a useful tool for research purposes in the domain of emotional speech analysis and recognition. An overview of 64 emotional speech data collections is presented in [3].

In this paper an acted emotional speech database was employed. Databases for emotional speech synthesis are usually based on acted speech, where some professional speakers read a set of texts simulating the desired emotions. The advantage of this method is the control of the verbal and phonetic content of speech as all the emotional states can be produced using the same phrases. This allows direct comparisons of the phonetics, of the prosody and of the voice quality for the different emotions. The disadvantage is the lack of authenticity of the expressed emotion.

One of the challenges is the identification of oral indicators (prosodic, spectral and voice quality) attributable to the emotional behavior. Many features for emotion recognition from speech have been explored, but there is still no agreement on a fixed set of features. The objective is to establish a relationship between the speaker's emotional state and some quantifiable parameters of speech. We present a data-mining experiment where we computed a set of acoustic features as wavelet energy and MFCC time series of the data.

Speech can be classified as information carrying non-stationary acoustical signals. Based on this, speech is a good candidate for wavelet analysis because its variability in style changes rapidly over time depending on the environment and on the speaker's characteristics.

In the next section, the emotional speech database and the wavelet transform are presented. Section 3 describes the feature extraction methods. Section 4 deals with the classification performance and discussion.

Finally, conclusions and future work are presented.

2 METHODS AND TOOLS

2.1 Database

In this paper, data was provided from the Berlin Database of Emotional Speech. The corpus contains 535 utterances, performed in 6 ordinary emotions and in neutral emotional state. Sentences are labeled as: happiness (71), anger (127), fear (69), boredom (81), sadness (62), disgust (46) and neutral (79).

The same texts were recorded in German by ten professional actors, 5 female and 5 male, this allowing studies over the whole group, comparisons between emotions and comparisons between speakers. The corpus consists of 10 utterances for each emotion type, 5 short and 5 longer sentences, varying between 1 and 7 seconds. A perception test with 20 subjects was carried out to ensure the emotional quality and naturalness of the utterances, and those more confused were eliminated [8].

2.2 Wavelet Transform

The commonly used tool for signal analysis is Fourier Transform, which breaks down a signal into constituent sinusoids of different frequencies. It has one major drawback: in frequency domain, information about time is lost.

This is not very important if the signal does not change much over time. The Fourier Transform is less useful in analyzing non-stationary signals, as speech, which has sharp transitions, drifts and trends. A time-frequency representation of signals can be performed using wavelets.

The Discrete Wavelet Transform (DWT) is computed by successive low-pass and high-pass filtering of the discrete time-domain signal as shown in Figure 1. At each level, the high pass filter produces detail information (D_i) while the low pass filter produces coarse approximations (A_i). The output of the filters is decimated in order to maintain orthogonality, halving the number of coefficients at each iteration. The approximations are filtered again at each decomposition step.

DWT is suitable for analyzing signals whose information is located in the low-frequency regions. However, the DWT may not be suitable for signals with the information mainly located in the middle- or high-frequency regions due to the wide bandwidth in the higher-frequency region.

While the DWT is implemented through iterative decomposition of approximation coefficients using a two-channel filterbank, the Wavelet Packet Transform (WPT) can be implemented through iterative decomposition of all coefficients, yielding an equal frequency bandwidth [5].

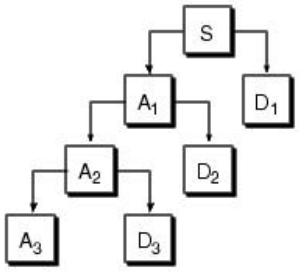

Figure 1: *The DWT as filter bank for a level 3 decomposition*

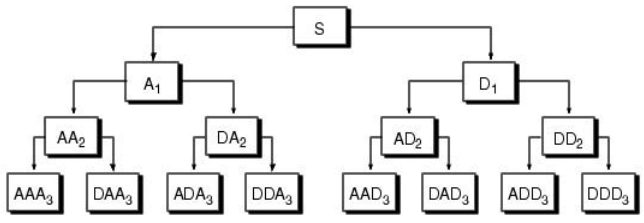

Figure 2: *WPT for a level 3 decomposition*

2.2.1 Perceptual Wavelet Packet Transform

In the proposed approach, the speech signal is also decomposed into critical subband signals by using the Perceptual Wavelet Packet Transform (PWPT) which is designed to match the psychoacoustic model. The sampling rate is of 16 kHz, yielding a speech bandwidth of 8 kHz. Within this bandwidth, there are 24 critical bands as shown in Figure 3. The decomposition is implemented by a 6 level tree structure; it integrates the concept of Mel scale and multi-resolution capabilities [3].

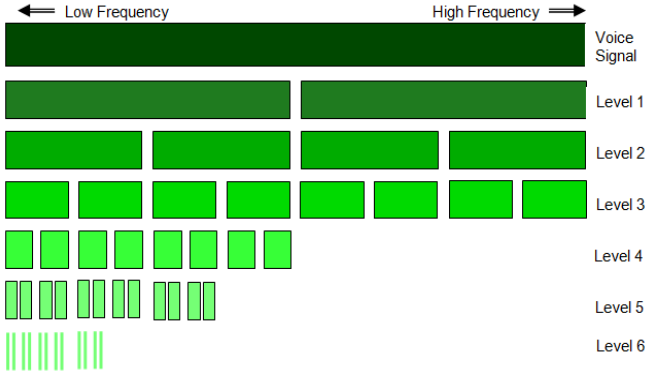

Figure 3: *Perceptual Wavelet Packet Transform*

3 FEATURE EXTRACTION

There are no established analytical methods in the field of voice analysis that can reliably determine the intended emotion carried by the speech signal. A possible approach in this context as seen in research is performing a trial to apply different and known signal processing methods, and to combine their results in such a way that there is a possibility for their pointing in the right direction - towards the emotion "hidden" in the signal.

The first purpose was to explore the spectral features by using the Mel-frequency cepstral coefficients (MFCCs). They have been widely employed in speech recognition also, due to superior performance when compared to other features. The Mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. For each speech frame of 30 ms, a set of Mel-frequency cepstrum coefficients was computed. The number of MFCC was chosen as 13. A series with the mean of MFCCs was detected for every utterance [4].

Wavelet Decomposition / No of iterations	DWT	WPT	PWPT
3	4	8	-
4	5	16	-
5	6	32	-
6	7	64	24

Table 2: *The number of wavelet energy coefficients*

Other, the speech signals are decomposed by using the-Discrete Wavelet Transform (DWT) respectively Wavelet Packet Transform (WPT) for 3, 4, 5 and 6 iterations. A wavelet subband decomposition is also used in a tree structure to divide the speech signal according to the Mel scale - Perceptual Wavelet Packet Transform (PWPT). For each sub band, the mean energy was calculated. Table 2 presents the number of the wavelet energy coefficients obtained for each situation.

Some additional features resulted from the time analyses of the speech waveform were also extracted:

- *relative mean square energy* – computed for each of the signal frames; the employed parameter is actually computed as a ratio between the maximum value of the energy frame in the waveform and the minimum value;
- *mean pitch frequency* – the pitch is computed for each of the appropriate frames, which provide maxima in the energy envelope, by using the autocorrelation method; the employed parameter is an averaged value of the pitch frequencies provided by each frame corresponding to the maxima in the mean signal energy;
- *normalized zero crossings rate* – the number of zero crossings is counted for each frame of speech signal and then the ratio between the largest and the smallest value is stored as a parameter.

4 EXPERIMENTS

In order not to favor one of the emotions over the others, the same number of utterances (45) was selected for every emotion during experiments. 20 utterances were used in the training step and the rest of 25 for testing.

The LibSVM tool developed by Chang and Lin was employed for classification task [9]. Support Vector Machines (SVMs) are a relatively new learning algorithms introduced by Vapnik (1995). They are based on the statistical learning theory of structural risk management. Support vector machines are built by mapping the training patterns into a higher dimensional feature space where the points can be separated by using a hyperplane. In LibSVM there are four kernels available: linear, polynomial, radial basis function and sigmoid. Optimal values for the kernel's parameters will be found by performing a grid search on the training data.

Experiments were made using different wavelet functions such as: Daubechies, Symlet, Coiflet and biorthogonal. Best results, achieved for Daubechies 4 are presented in Figure 4.

The most suitable wavelet decomposition seems to be DWT, 4 iterations. The overall accuracy in this case is 95.42% for a small feature vector – 18 coefficients (13 MFCC + 5 DWT). The used SVM kernel was the polynomial one.

Although Perceptual Wavelet Packet Transform seems to be very proper for others area of application connected to speech processing, for speaker emotional state identification the Discrete Wavelet Transform offers the best solution.

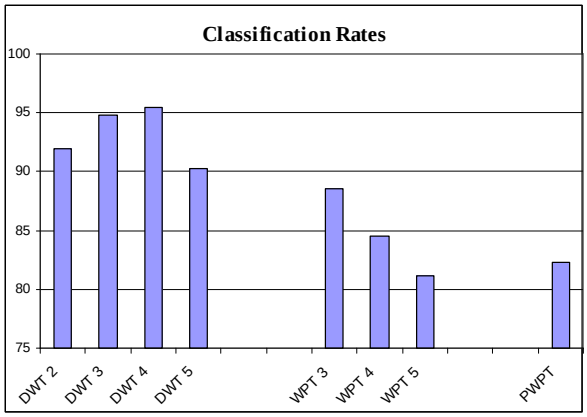

Figure 4: *Emotions Recognition Rates*

By adding the time analyses features, the classification rate is 96.57% for a 21 coefficients feature vector.

If we consider the entire database and we use 20 utterances for training and the rest for testing, the resulted accuracy is 83%. The used classifier was SVM, polynomial kernel and the feature vector was formed by 13 MFCC + 5 DWT + 3 time analysis coefficients.

5 CONCLUSIONS

The presented article has been focused on feature extraction methods that are useful in emotion recognition from speech signal. The extracted features were estimated over the whole utterance. This choice was made, due to the fact that in the literature global statistics is generally thought to be more suitable for this area of applications.

In future, we intend to introduce, other voice parameterization in order to minimize the confusion between states. We also plan to investigate other wavelet techniques that can be used to overcome some of the deficiencies in the methods presented.

REFERENCES

[1] J. Sidorova, "Speech Emotion Recognition" DEA report, Universitat Pompeu Fabra, July 4, 2007

[2] A. Batliner, "Whodunnit - Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech", *Journal: Computer Speech and Language*, Volume 25 Issue 1, January

[3] O. Farooq, S. Datta, "Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition", IEEE Signal Processing Letters, ISSN: 1070-9908, 2001, pp. 196 – 198

[4] R. Hasan, M Jamil, G Rabbani, "Speaker Identification using Mel Frequency Cepstral Coefficients', International Conference on Electrical & Computer Engineering, ICECE 2004

[5] S. Mallat, *A wavelet tour of signal processing*, New York, Academic Press, 1999

[6] D.Ververidis, C. Kotropoulos, „Emotional speech recognition: Resources, features, and methods", Speech Communication, vol 48, 2006, p 1162–1181

[7] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," IEEE Transaction on Speech and Audio Processing, vol. 8, no. 4, pp. 429–442, Jul 2000.

[8] http://pascal.kgw.tu-berlin.de/emodb/index-1280.html

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm/