

GRADIENT BASED SPECTRAL PEAK LOCATION FOR NOISE ROBUST SPEECH RECOGNITION

Penny Hix
Old Dominion University
phix@odu.edu

ABSTRACT

In this paper a gradient-based algorithm for finding spectral peak locations is presented. The algorithm makes use of gradient and acceleration locations in the spectrogram for locating the peaks. Use of frequency gradients and accelerations locate peaks. The results are then interpolated to yield a smooth peak envelope. The method is evaluated in the aurora framework. A first pass locates all spectral peaks and automatically eliminates low magnitude, high frequency peaks that are likely to contain more noise than speech information. The second pass widens the spectral peaks with spectral information. This widening is expected to increase automatic speech recognition based on the peak envelope discrete cosine transformation feature representation.

1. INTRODUCTION

Recent research has demonstrated that in speech signals with low SNR the regions surrounding spectral peaks contain higher levels of information relevant to feature extraction and subsequent automatic speech recognition (ASR) [1]. The method presented is a simple frequency based dynamic programming algorithm that utilizes the spectral slope and acceleration to locate peak candidates. Other methods

typically use the standard mel-frequency cepstral coefficients (MFCC). Ours differs in the combination of resulting peak envelope with the computation of subsequent features by utilizing discrete cosine transform coefficients (DCTC) to obtain the final feature vectors. In previous work [1] it has been established that the DCTC yields similar spectral results with less computational demand.

The slope-acceleration peak envelope based features are computed by parameterization of the information surrounding the spectral peaks. The relatively high SNR speech information captured in the peak regions is expected to yield improved robustness in noisy conditions.

This paper is organized as follows: Section 2 provides a basic description of the computation of the spectrum. Sections 3 and 4 give a short introduction to the peak location with widening and peak envelope based feature computation, respectively. Section 5 explains the experimental setup for evaluating the features obtained using the peak envelope spectrum. Section 6 presents the experimental results.

2. SPECTRUM COMPUTATION

Speech is normally acquired through the use of a microphone. The resulting speech signal is in analog form. Thus, the first step in computing the spectrum is converting the speech signal from

analog to digital form. Figure 1 shows the subsequent steps involved in feature extraction. The digital speech signal is then filtered to emphasize the center of the signal. The pre-emphasized signal is windowed with a Kaiser window with beta 6. A 512-point Fourier transform (FFT) is then taken. The magnitude of the FFT provides our spectral representation of the original speech signal. Subsequent steps in the diagram yield the dynamic features of the speech signal. The peak location and smoothing in this paper are inserted immediately following the log magnitude step. The peak location index vector is then used in the computation of the Discrete Cosine Transform Coefficients (DCTC). The resulting static features are processed via the discrete cosine transform (DCS) to obtain the final set of dynamic features.

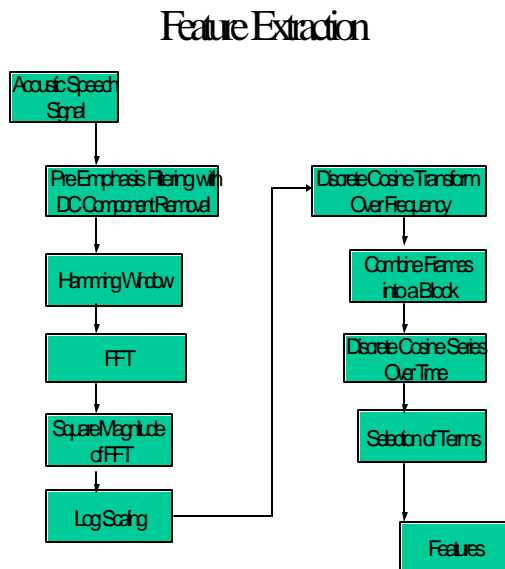


Figure 1

3. LOCATING PEAKS

The gradient-acceleration method is motivated by recent work utilizing the

YAAPT algorithm [3] to identify peak candidates. Peak candidate locations obtained via the YAAPT algorithm were then used to determine potential pitch cycle markers [4]. Spurious, or erroneous, peak candidates can lead to incorrect pitch cycle markers; thus, the need to achieve improved peak location in speech signals.

An important aspect of this approach is the fact that no spectral markers are required for identification of peak candidates. This mitigates possible inclusion of incorrect peak locations. Inclusion erroneous of peak candidates would yield false speech information in the parameterization of the speech features.

The gradient, of the spectrum obtained in section 2, is taken, $Y = grad(X)$. Taking the gradient of Y, we obtain the acceleration vector, $Y' = grad(Y)$. The gradient, Y, and acceleration, Y' , are used together to locate peaks. The locations of the peaks are saved in an index vector, X. The peaks and index vectors are used to locate spectral information immediately surrounding peaks.

As a result of the peak location activity we also obtain the frequency index for peak candidate locations. This frequency index is used in obtaining spectral information surrounding the peak candidates. The additional points are used to widen, or increase, the spectral information surrounding peak candidates.

4. PEAK ENVELOPE FEATURE COMPUTATION

Vectors obtained via peak location and widening typically vary in length. Whereas, automatic speech recognition systems normally require feature vectors

of uniform length. This problem is addressed by using the frequency index vector to compute static coefficients via the discrete cosine transform. The frequency index vector points to the peak candidates as well as the spectral information immediately surrounding them. Also, it is important to note that one frequency index vector is computed for each frame of speech. Thus, the resulting matrix of indices is quite sparse.

The computation of the peak based discrete cosine transform coefficients provides enhanced spectral peaks and smoother spectral valleys. However, the gradient-acceleration based peak locations are each computed for a single time frame. This approach may provide incorrect variations in peak locations from one time frame to the next. For this reason a temporal constraint may be required in order to provide more consistent peak locations.

5. EXPERIMENTAL SETUP

The Aurora 2 and Aurora 3 databases are used for experiments. These databases contain connected digit strings spoken in clean and in noisy environments. Additive and convolution (channel) noise are both provided in the Aurora databases. In the noisy speech Signal-to-Noise Ratio (SNR) levels vary from SNR -5 to SNR 20. The speech recognition is performed using Hidden Markov Models computed with the Hidden Markov Model Toolkit (HTK). Peak envelope based features are compared with results obtained by the Aurora standard front-end, and also with standard HTK front-end features. Aurora based features are computed using mel-frequency cepstral coefficients computed. The resulting dynamic feature vectors contain 39

components, 13 static coefficients, 13 delta coefficients, and 13 delta-delta coefficients.

Features computed using HTK are also MFCC based. For each set of features the same HTK setup is used for HMM model computation and subsequent Viterbi decoding. The HMMs are 18 state, three mixture, tied models.

6. EXPERIMENTAL RESULTS

The HMM based recognizer is trained on 8440 speech samples. The primary factor that affects performance is the width of the peaks. Figure 2 shows the spectrum (in blue) and the peak information (in red) for the utterance 1978213.

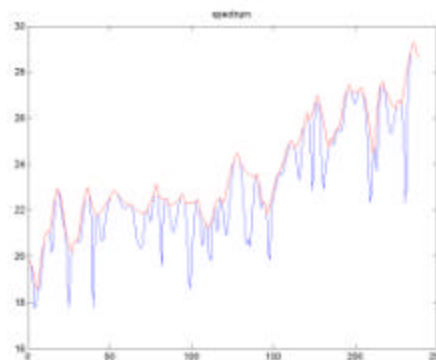


Figure 2

Table 1 gives results of experiments conducted for evaluation and comparison of automatic speech recognition performance utilizing features computed with the peak envelope method. Average results are given for clean speech as well as several noise levels.

# features	39			
Aurora	78.8			
HTK	78.1			

ODU(peaks)	81.9			
------------	------	--	--	--

From the table, it can be observed that the peak envelope based method is more robust to noisy conditions. In noisy conditions this method is comparable the standard method using MFCC features.

Unfortunately, the method does not improve performance on clean speech. One reason for this phenomenon is that the peak envelope method ignores important information contained in non-peak areas of clean speech. One possible solution to this is to combine the peak envelope method with standard feature computation for clean speech conditions. This approach may provide robust performance in a wider range of conditions.

The table gives results for HMM, Aurora, and ODU peak envelope features. For each set of features all other experimental settings are the same. These results show that the peak envelope features provide improved accuracy for most conditions.

6. CONCLUSION

A gradient-acceleration based method to locate spectral peaks for subsequent feature computation has been presented. This method mitigates the inclusion of spurious peaks, leading to feature vectors that contain erroneous speech information. Use of gradient-acceleration located peaks yields more reliable information because spectral peaks are enhanced and spectral valleys are smoothed. Experimental results show robustness to high SNR levels.

Further work will be to use properties of the sparse index matrix to speed up computations, add temporal constraints

to achieve more consistent peak location, and to combine peak envelope (for noisy speech) method with standard DCTC method (for clean speech) to obtain robust performance in a wider range of conditions.

7. REFERENCES

- [1] Shajith Ikbal, Herve Boursard, Mathew Magimai, "HMM/ANN Based Spectral Peak Location Estimation for Noise Robust Speech Recognition," in Proc. Of ICASSP-05, Philadelphia, March 2005, I452-I456
- [2] Lurng-Kuo Liu, Feig, E. , "A block-based gradient descent search algorithm for block motion estimation in video coding", in IEEE Transactions on Circuits and Systems for Video Technology, 419-422, August 1996
- [3] Kavita Kasi, Stephen Zahorian, "Yet Another Algorithm for Pitch Tracking ", in Proc. ICASSP-02, Orlando, May 2002
- [4] Princy Dikshit, Stephen A. Zahorian, Shivram Nagulapati, "An Algorithm for Locating Fundamental Frequency Markers in Speech Signals," in Proc of ICASSP-05, Philadelphia, March 2005, I233-I236