ANALYSIS OF SPEECH SEGMENTS USING VARIABLE SPECTRAL/TEMPORAL RESOLUTION

Xihong Wang, Stephen A. Zahorian, Stefan Auberg

Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA

ABSTRACT

In this paper we present an approach for efficiently computing a compact temporal/spectral feature set for representing a segment of speech, with effective resolution depending on both frequency and time position within the segment. The goal is to mimic the resolution properties of the human auditory system, but using a computationally efficient FFT-based front end rather than a more complex auditory model. In particular we apply both frequency and time "warping" to FFT spectra to obtain good frequency resolution at low frequencies and good time resolution at high frequencies. Time resolution is also varied so that the center of the segment is better represented than the endpoints. The resolution can be varied by the selection of "warping" functions controlled using a small number of parameters. The method was experimentally verified for the classification of six stops extracted from the TIMIT continuous speech data base. The best classification rate obtained was 81.2% for test data using 50 features computed with the method presented.

1. INTRODUCTION

Normally spectral analysis of speech signals begins with a fixed frame-rate short-time Fast Fourier Transform (FFT) calculation which inherently has uniform time and frequency resolution. Tradeoffs between time and frequency resolution can be made by varying the frame length and frame spacing. Typical values are a frame length of 20 ms (Hamming window) and a 10ms frame spacing. In order to approximate the frequency resolution properties of human hearing, the FFT spectrum is sometimes resampled according to a mel or Bark scale frequency "warping" function. Alternatively, FFT values are summed to form an auditory scale based filter bank. Although these methods do provide a good approximation to auditory frequency scales, they do so primarily via additional smoothing at higher frequencies, and do not take advantage of the higher time resolution which is then theoretically possible at these higher frequencies. Approaches to take advantage of variable time-frequency resolution include auditory models and wavelet analysis. Although these two methods (particularly auditory models) undoubtedly better match human hearing sensitivity both temporally and spectrally, they have not yet been shown to be clearly superior to FFT-based methods for automatic speech recognition and are not widely used. Problems include computational complexity and difficulties in extracting a small number of features from the initial analysis. In this paper we describe an FFT-based method for speech spectral analysis, which begins with an "over-sampled FFT," and converts the results of the FFT magnitude spectrum to a small number of features which represent the spectrum with frequency resolution approximating an auditory frequency scale, and time resolution which is better at high frequencies than low frequencies. Implicit in the method is the use of speech segments, each consisting of several frames spanning an interval containing most of the acoustic information for a single phone, for feature calculations.

2. PRINCIPLE

The principle of this method is that an original log spectrum, $X(t, f)$, computed with uniform (and high) time/frequency resolution over the duration of a speech segment, is transformed to a "perceptual" time t' and frequency f' domain, to achieve the desired time-frequency resolution. The transformed frequency f' depends on physical frequency f only, but the transformed time t' depends on both t and f. These transformations can be represented by a time-independent frequency "warping" function and frequency-dependent time "warping" function. The methods for computing features from this perceptual time/frequency spectrum are described in this section.

Let $X(t, f)$ denote the original log spectrum, computed with a uniform time/frequency resolution, where, for convenience, t and f both are normalized to the range [0,1]. Also define "perceptual" time and frequency variables t' and f' in the interval [0,1] which are related to t and f by

$$t' = h(t, f) \qquad \text{and} \qquad f' = g(f).$$

In this equation, g is a typical warping function over frequency, such as Bark or mel function. Notice that h, the time warping function, depends on both time and frequency. The motivation for h is to enable variable resolution as function of both position within a segment and frequency. h is typically selected to achieve higher resolution at the middle of a segment relative to the beginning and endpoints and also better resolution at high frequencies than at low frequencies. The "perceptual" spectrum X' , is then the transformed version of the original spectrum X, which incorporate the g and h warping functions, according to

$$X'(t', f') = X(t, f).$$

In the perceptual frequency (f') domain, if *more* details are preserved in one interval than in another interval, then we say that the frequency resolution is higher in that interval. In terms of the

relationship between *f'* and *f*, the resolution in *f'* is proportional to | dg/d*f* |. Thus if *f'* is sampled uniformly, higher resolution with respect to *f* is obtained for those intervals with a large value of dg/d*f* , or close sampling with respect to *f*. For example a Bark warping function has much higher derivatives at low frequencies than high frequencies. Similarly, the resolution of *t'* is proportional to | d*h*/d*t* |. Thus time and frequency resolution can be controlled by the choice of *h* and *g* functions. For speech processing high frequency resolution is desired at low frequencies, while high time resolution is desired (or at least possible) at high frequencies. There are of course limits on the "real" time and frequency resolution in *t'* and *f'*, dependent on the initial resolutions in *f* and *t*. Typically *g* is an approximation to a Bark function, such as a bilinear warping function with a warping coefficient of about .5. Typically *h* is a function with a smooth derivative, symmetric about the center and with highest values in the center of the segment. In this paper a family of integrated Kaiser windows was used for *h*, with a different window at each frequency . That is, for each *f*, dh/dt was a Kaiser window. By adjusting each Kaiser window's β, with highest β 's at highest frequencies, we obtained better time resolution at the middle of a segment and at high frequencies.

A very important aspect of the material presented in this section is that acoustic features *Feat(i, j)* for compactly representing spectral/temporal information can be computed using a cosine expansion of *X'* over *t'* and *f'*, using

$$Feat(i,j) = \int_{t'=0}^{1} \int_{f'=0}^{1} X'(t',f') \cos(j\pi t') \cos(i\pi f')\, df'\, dt' \qquad (1)$$

With a change of variables, Equation 1 can be rewritten in terms of linear *t* and *f*, and with the functions *g* and *h*, as

$$Feat(i,j) = \int_{t=0}^{1} \int_{f=0}^{1} X(t,f) \cos(j\pi * h(t,f)) \cos(i\pi * g(f)) \frac{\partial g}{\partial f} df \frac{\partial h}{\partial t} dt \qquad (2)$$

Equation 2 can be more conveniently interpreted by defining frequency-dependent basis vectors over time as

$$\Phi j(t,f) = \cos(j\pi * h(t,f)) \frac{\partial h}{\partial t} \qquad (3)$$

and modified basis vectors over frequency as

$$\Theta i(f) = \cos(i\pi * g(f)) \frac{\partial g}{\partial f} \qquad (4)$$

Using the basis vectors defined in Equations 3 and 4, and by rearranging terms, Equation 2 can be rewritten as

$$Feat(i,j) = \int_{f=0}^{1} \Theta i(f) \int_{t=0}^{1} X(t,f) * \Phi j(t,f)\, dt\, df \qquad (5)$$

3. IMPLEMENTATION ISSUES

In order to compute features for speech analysis as per Eq. 5, the log magnitude spectrum is originally computed using an FFT for each of several frames over a segment. For the experiments reported in the next section, a Hamming-windowed frame length of 10 ms and a frame spacing of 2 ms were used, providing an initial frequency resolution on the order of 100 Hz and time resolution on the order of 10 ms. These values thus represent the limiting resolutions possible in subsequent representations. All integrations above were replaced with sums. For computational efficiency, the basis vectors were precomputed and stored. We refer to the basis vectors over time $\Phi_j(t, f)$ as Discrete Cosine Series (DCS) basis vectors and the basis vectors over frequency $\Theta_i(f)$ as Discrete Cosine Transform (DCT) basis vectors. Unlike most spectral/temporal speech processing methods, the calculations over time were done first, and the calculations over frequency secondly. Also note that several sets of basis vectors over time were used, one set for each frequency.

For typical analysis conditions the total computations required for these features are slightly less than if an expansion is performed over frequency first, and then over time. The computations are much less than those required using a two-dimensional non-partitioned basis vectors set . For example, assuming a 300 ms segment, 150 total frames, and 200 frequency samples the total calculations (in terms of multiply-adds needed after the original spectrum is computed) to compute 50 features (10 DCTs and 5 DCSs) are 160,000 for the method presented in this paper, versus 307,500 if sums over frequency are computed first, versus 1,500,000 if 2-D basis vectors are used (Silsbee, et al., 1994).

The features *Feat(i, j)* computed with Equation 5 effectively smooth the original spectrum, in the sense that a smoothed version of *X* can be reconstructed from *Feat(i, j)* . However, the degree of smoothing depends on position in time and frequency, as specified by *g* and *h*. A series of time/frequency plots are shown in Fig. 1 to illustrate this, using an 80 ms section of speech signal centered at the burst of a /d/ extracted from a sentence. Panel 1 is the original log spectrum computed using a series of FFT spectra (40 frames, 5 ms frame length, 2 ms frame spacing). Panel 2 is the smoothed spectrum reconstructed from the data of panel 1, using 10 basis vectors over frequency (DCTs) and 5 basis vectors over time (DCSs) without any warping (50 total "features"). Thus, on the average, panel 2 illustrates good frequency resolution and poor time resolution, since it was derived from a large number of DCTs and a small number of DCSs. Panel 3 is also a smoothed spectrum computed from 50 features, but with 5 DCT basis vectors and 10 DCS basis vectors. Therefore panel 3 contains more time information and less frequency information than does panel 2. Panel 4 is the spectrum smoothed by the method described in this paper, with the Kaiser window time warping varying from 0 at low frequencies to 6 at high frequencies. Thus, although ten DCT terms and five DCS terms were still used, the low frequency spectrum contains more frequency resolution, whereas the high frequency spectrum has more temporal resolution.

4. EXPERIMENTS

In order to evaluate the method described in this paper, several experiments were conducted for the speaker and context independent classification of the six stops /b, d, g, p, t, k /. The SX sentences from the TIMIT database were used. A small number of experiments were also performed using the telephone version of TIMIT (NTIMIT). The analysis was based on a specified segment length centered at the onset of the stop burst. For the initial experiments which were used for parameter tuning

and general refinement of the method, 450 speakers were used for training and 49 speakers for testing. For the experiments with numerical results reported in this paper, the training set consisted of 499 speakers (357 male, 142 female) for a total of 2495 sentences. The training set contained 1384, 1388, 811, 1699, 2365, 2412 phone tokens respectively for the 6 stops in the order mentioned above. The test set contained 50 speakers (33 male, 17 female), or 250 total sentences, and 141, 134, 77, 155, 205, 246 phone tokens. This particular speaker arrangement was the same as that used in some previously reported work (Goldenthal and Glass , 1993).

The classifier used in these experiments was a binary-pair partitioned (BPP) neural network (Zahorian et al, 1993) . For the BPP classifier, an individual network is trained for each pair of categories, with final decisions based on the output of all classifiers. For the results reported in this paper a total of 15 pairwise networks were used for the six stops. Each network was a fully interconnected feed-forward perceptron structure with unipolar sigmoidal nonlinearities, with one hidden layer (25 nodes), and a single output node. Each network was trained with backpropagation for 100,000 network updates, beginning with an initial learning rate of .25, reduced by a factor of .96 after every 5,000 iterations. In our previous work with vowels, we achieved slightly higher rates with this classifier as compared to a single large network.

For the primary experiments reported in this paper, the initial spectrum was computed with a frame length of 10 ms (after multiplication by a Hamming window) and a frame spacing of 2 ms, using a 512 point FFT. The frequency range used for feature calculations was 100 Hz to 6000 Hz. Frequency warping was accomplished with bilinear warping, using a coefficient of .45. Time "warping" was accomplished such that the derivative of the *h* function mentioned above was a Kaiser window with different β-values controlled by two parameters: *Time warp begin* and *Time warp end*. The time warping factor was then linearly scaled according to frequency position between these beginning and ending values.

Numerous preliminary experiments were done with the 450 speaker training set to determine the general performance of this method and to determine "good" values for several of the arbitrary parameters. The objective was also to determine which options or parameter settings had very little effect on classification accuracy. For example, the use of a segment window to first multiply all frames in the original spectrum over time using the first time basis vector, (as was used to generate Figure 3 had almost no effect. Similarly, results obtained with and without orthonormalization of the basis vectors were nearly identical. Therefore neither the window multiplication nor orthonormalization were used in the primary experiments. Results based on an FFT length of 512 were found to be slightly superior to those obtained with a 256 point FFT, thus motivating the use of the 512 point FFT. The frame length of 10 ms and frame spacing of 2 ms was found to be somewhat better than either 5 ms frame length (and 1 ms spacing) or 10 ms length (1 ms spacing), thus motivating the use of the 10 ms frames. The frequency warping factor was varied from 0.25 to 0.75 and the biggest difference was only 0.8%. Windows other than the Kaiser window were investigated to control the time warping. In general the Kaiser window was slightly better than the others tried. Based on these preliminary experiments we chose and fixed some of the parameters which we thought would give the "best" results. In the next few paragraphs we describe and summarize results for some of these experiments, all conducted with the data mentioned above.

4.1. Experiment 1

This experiment was conducted to determine the effects of segment length on classification rate for the six stops. The general conditions were as previously mentioned. For each case 50 features were computed (10 DCT terms and 5 DCS terms) . The segment length was varied from 25 ms (13 frames) to 300 ms (150 frames). For each case, various values for the beginning and end values of the time warping parameters were used. The test results for three cases are given in Table 1 – time warp of 0:0 (no time warping), time warp of 10:10 (best fixed time warping), and time warp of (5:30) (best variable resolution time warping). For the shortest segments, the best results was obtained with no time warping. The best overall results were obtained with the longest segment length of 300 ms. The overall best result of 81.2% was obtained with 300 ms and the variable resolution warping.

Seg. Length	25 ms	50 ms	100 ms	150 ms	200 ms	300 ms
time warp 0:0	62.5	73.8	78.3	78.8	79.3	75.9
time warp 10:10	56.2	69.2	75.6	79.0	78.6	80.8
time warp 5:30	54.2	68.3	74.8	78.5	78.5	81.2

Table 1: Classification rates for six stops for various segment lengths and various time warping conditions.

4.2. Experiment 2

The first experiment indicated that the best results are possible with a long segment. In this experiment, focusing on the 300 ms segment, we examined in more detail the classification results for several different values of beginning and ending values of the time warping parameter. Results are presented in Table 2. The results vary by only a small amount for all conditions tested, except for no time warping. However, three of the variable warping conditions are better than the best fixed warping condition.

a. Fixed time warping:

time-warp	0:0	5:5	10:10	15:15
classification rate	75.9	79.4	80.8	78.7

b. Frequency-dependent time warping:

time-warp	3:30	3:45	5:15	5:30	5:45

classification rate	80.9	80.4	80.6	81.2	81.0

Table 2: Classification rates for six stops using a 300 ms segment and various warping conditions.

5. CONCLUSION

A variable time/frequency resolution method for computing a set of acoustic speech features was presented along with experimental results for the classification of stop consonants. Several experimental conditions were tested with different numbers of features, different segment lengths, different time warping functions, and different amounts of time warping. The best variable time warping was better than the best fixed time warping by a small amount. Substantially higher results are obtained using long segments (300 ms) versus short segments (25 ms). The best test results (81.2% for regular TIMIT and 68.2% for NTIMIT) are, to our knowledge, the best reported results for stops from TIMIT. These results were also obtained with a relatively small number of features (50) and with less computational complexity than is usually required with an auditory model implementation. In general this method emphasizes the need for and enhances segment-based speech feature processing relative to frame-based analysis.

Xihong some points

Do all references match up using dates?
Do we mention the frequency warping value we actually used for main tests ?
Are we using italics for all g and h in equations?

Do we actually reference no 5?

Equations i,j ordering should be consistent.

Are there any other important conditions for experiments that we accidentally removed?

6. REFERENCES

1. Goldenthal, W. D. and Glass, James R., "Modeling Spectral Dynamics for Vowel Classification", *Euro Speech 93 , vol*. 1 289-292, 1993.

2. Nossair, Z., Silsbee, P., and Zahorian, S., " Signal Modeling Enhancements for Automatic Speech Recognition", *ICASSP-95. vol. 1,* 824-827, 1995.

3. Silsbee, P., Zahorian, S., and Nossair, Z., " A Warped Time-frequency Expansion for Speech Signal Representation", *Proc. IEEE-SP Symp. on Time-Frequency and Time-Scale Anal:* 636-639,1994.

4. Zahorian, S., Nossair, Z., and Norton, C., " A Partitioned Neural network Approach for Vowel Classification Using Smoothed Time/Frequency Features", *Euro Speech 93 , vol*. 2 1225-1228, 1993.

5. Zavaliagkos, G., Zhao, Y., Schwartz, R., and Makhoul, J., "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, *vol*. 2, *no*. 1, 1994, 151-160.