INTERSPEECH 2009 BRIGHTON

Spectral and Temporal Modulation Features for Phonetic Recognition

Stephen A. Zahorian, Hongbing Hu, Zhengqing Chen, Jiang Wu

Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902, USA

{zahorian, hongbing.hu, zhengqing.chen, jiang.wu}@binghamton.edu

Abstract

Recently, the modulation spectrum has been proposed and found to be a useful source of speech information. The modulation spectrum represents longer term variations in the spectrum and thus implicitly requires features extracted from much longer speech segments compared to MFCCs and their delta terms. In this paper, a Discrete Cosine Transform (DCT) analysis of the log magnitude spectrum combined with a Discrete Cosine Series (DCS) expansion of DCT coefficients over time is proposed as a method for capturing both the spectral and modulation information. These DCT/DCS features can be computed so as to emphasize frequency resolution or time resolution or a combination of the two factors. Several variations of the DCT/DCS features were evaluated with phonetic recognition experiments using TIMIT and its telephone version (NTIMIT). Best results obtained with a combined feature set are 73.85% for TIMIT and 62.5% for NTIMIT. The modulation features are shown to be far more important than the spectral features for automatic speech recognition and far more noise robust.

Index Terms: modulation spectrum, spectral-temporal feature, discrete cosine transformation, phonetic recognition

1. Introduction

The techniques for extracting useful information from short-time speech spectral analysis have been developed from research in speech signal processing and Automatic Speech Recognition (ASR) over the past several decades. The prevailing view for many years was that “static” spectral information, obtained from frame-based features such as Mel Frequency Cepstral Coefficients (MFCCs) or formants, are the most critical features for ASR. In contrast, “dynamic” features, such as the delta coefficients for MFCC, representing spectral trajectory information over short time intervals of approximately 50 ms, were considered to be of secondary importance. More recently, dynamic features extracted from temporal trajectories of spectral information over longer time intervals have been shown to be very effective features for noise robust ASR.

Beginning in the mid-1990s, Hermansky and Morgan [1] pointed out the importance of temporal trajectory information by introducing RelAtive SpecTrA (RASTA) features. In the works of [2, 3], the modulation spectrum has been demonstrated to be a very useful speech representation for ASR. For instance, Kanedera et al. [3] combined bandpass filtering with RASTA processing and extracted different modulation spectrum components from different frequency bands. Kingsbury et al. [4] used a modulation spectrum as a front end in speech recognition and obtained significant improvement by combining it with log-RASTA-Perceptual Linear Predictive coding. In more recent research, in order to achieve consistently low word error rates, Valente and Hermansky [5] proposed hierarchical and parallel processing techniques combining outputs of independent classifiers and modulation frequency channels. In addition, the use of the auditory modulation spectrum was also investigated in greater depth in their work.

From a conceptual point of view, modulation features are most typically computed by dividing the log spectrogram into frequency bands or “slits “ and then computing features which reflect the temporal evolution of the spectra in each band, with each frequency band analyzed separately. These modulation features are then usually combined with spectral features such as MFCCs or Perceptual Linear Prediction (PLP) coefficients. Since the combination of modulation features and spectral features results in very large feature vectors (on the order of 150 terms), researchers have studied the best methods for combining and using all these features, such as the work reported in [5].

In this paper, we propose a quite different approach, whereby modulation features are considered as a Discrete Cosine Series (DCS) analysis over time of the Discrete Cosine Transform (DCT) coefficients of the log magnitude spectrum. Thus, rather than each modulation spectral term being associated with a separate spectrogram band, each modulation term is computed from an integrated feature of the entire frequency spectrum. The DCT/DCS feature set combines the spectral and temporal modulation features. However, the feature set can be tuned to emphasize spectral information, or modulation information, or a combination of the two types of information.

2. Approach

More details of the DCT Coefficients (DCTCs) and DCS Coefficients (DCSCs) features are given in the works of Zahorian et al. [6], and Kajanadecha and Zahorian [7]. Summarizing briefly, the basic idea is to compute frame-based spectral features as a modified cosine transform of the spectrum and then to compute feature trajectories with another cosine transform over time.

The first step of this feature calculation is to compute DCTC terms from the spectrum X, with the frequency f normalized to a [0, 1] range, as follows

$$DCTC(i) = \int_0^1 a(X(g(f)))\Phi_i(f)df \,. \quad (1)$$

In this equation, i is the DCTC index, $a(X)$ is a nonlinear amplitude scaling and $g(f)$ a nonlinear frequency warping. $\Phi_i(f)$ is the ith basis vector over frequency computed as

$$\Phi_i(f) = \cos[\pi i g(f)]\frac{dg}{df} \,. \quad (2)$$

The crucial elements of this approach are the selection of the nonlinear amplitude scaling $a(X)$ and the nonlinear frequency scaling $g(f)$, so that the cosine transform is with respect to a "perceptual" scale. In practice, the scaling $a(X)$ is typical a log, and the scaling $g(f)$ is a Mel-like function.

The spectrum is computed with an FFT computed from overlapping windowed speech frames and the integral is computed using a sum over a selected frequency range. The frequency warping function is implemented both through the modified cosine basis vectors and through the use of interpolated spectral magnitude values obtained from the FFT.

To illustrate the process, the first 3 DCTC basis vectors, using an approximation to Mel warping, are shown in Figure 1. For the results reported in this paper, the warping function is given by

$$f' = 2.0959 \times \log_{10}(1 + \frac{f}{0.5}). \quad (3)$$

In this equation, f' is considered perceptual frequency and f is frequency in Hz, except that both f and f' are normalized over [0, 1].

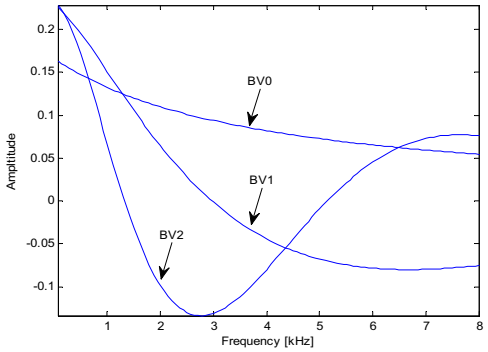

Figure 1: *First 3 DCTC basis vectors*.

In order to create the DCSC features that represent the spectral evolution of DCTCs over time, thus encoding the modulation spectrum, a cosine basis vector expansion over time is performed using overlapping blocks of DCTCs so that the temporal resolution is higher in the central region than for the end regions. That is, the DCSCs are computed as

$$DCSC(i, j) = \int_0^1 DCTC(i, h(t))\Theta_j(t)dt, \quad (4)$$

where $\Theta_j(t)$ is the jth basis vector over time computed as

$$\Theta_j(t) = \cos[\pi i h(t)]\frac{df}{dt}. \quad (5)$$

In this equation, $h(t)$ is a time warping function and t is normalized to [0,1] over a selected segment (a "block"). In practice, t is discrete, corresponding to a frame index, and the integral is computed using a sum of all frames in the block. The calculation is repeated for each overlapping block, with the block spacing some integer multiple of the frame spacing.

A set of 3 typical DCSC basis vectors is shown in Figure 2. Using these modified basis vectors, feature trajectories are represented using the static feature values for each frame, but with varying resolution over a block consisting of several frames. $DCSC(i,j)$ are the set of spectral-temporal features that represent speech for a block of frames.

Several parameters in the DCTC/DCSC analysis can easily be varied to examine tradeoffs between static spectral information and trajectory spectral information in terms of effects on robustness of ASR systems. For instance, for increased emphasis of spectral information, the frame length used to compute the spectra can be relatively long (on the order of 25 ms), and the number of DCTCs can be large (on the order of 15 to 20 coefficients). To evaluate the effectiveness of purely static spectral information, the DCSC step can be eliminated and the ASR system tested with DCTCs directly. For increased emphasis on trajectory information, a short frame length and frame spacing (on the order of 5ms and 1ms respectively) can be used along with a large number of DCSC terms (on the order of 5 to 10).

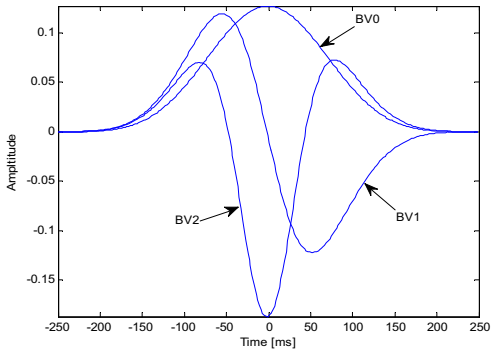

Figure 2: *First 3 DCSC basis vectors.*

In the following experimental section, features that emphasize frequency resolution are referred to as spectral features and features that emphasize the temporal trajectory of the highly smoothed spectrum are referred to as modulation spectrum features.

3. Experimental Evaluation

Experiments were conducted to evaluate the proposed method using the TIMIT database. The SA sentences were removed from the database, resulting in 3696 sentences from 462 speakers for training and 1344 sentences from 168 speakers for test. Experiments were also conducted with the telephone version of TIMIT, i.e. NTIMIT. All analysis parameters were identical for the TIMIT and NTIMIT evaluations, except for frequency range. For TIMIT, the frequency range was selected to be 50 to 7000 Hz, whereas for NTIMIT the frequency range for analysis was set at 300 to 3400 Hz.

The signal to noise ratio (SNR) was varied from no added noise (clean) to 0 dB, in steps of 10 dB (totally 5 noise conditions). A reduced 39 phone set as used in [8] was mapped down from the original TIMIT 62 phone set and used in the experiments.

Left-to-right 3-state Hidden Markov models with no skip were used and a total of 48 (eventually reduced to 39 phones) context independent monophone HMMs were created from the training data using the HTK toolbox (Ver3.4). The bigram phone information extracted from the training data was used as the language model. For all experiments with DCTC/DCSC features, a frame spacing of 8 ms (125 frames per second) was used. In an attempt to extract the most possible information from the speech features tested, a large number of mixtures were used to model each state with a diagonal covariance matrix. The actual number of mixtures was varied from 25 to 75, as mentioned for each evaluation. The number of mixtures was increased for the higher dimensionality feature sets and lowered for the lower dimensionality feature sets, since this approach yielded highest accuracies with stable HMM models.

The objective of the experiments was to compare phoneme recognition accuracy of control features (13 MFCCs with delta and acceleration terms, or 39 total terms) with static spectral features, with primarily modulation spectrum features, and with a combination of spectral and modulation features. More details are given for each experimental condition.

3.1. Experiment 1: Control

The intent of this experiment was to establish a baseline for ASR phoneme accuracy using "conventional" features, and the identical HTK recognizer and database configuration as was used for the proposed features.

The MFCC features were computed directly with the HTK supplied front end, using the typical parameter settings for MFCCs as well as delta and acceleration terms. That is, 12 MFCCs plus energy with delta and acceleration terms, or 39 total terms were obtained every 10 ms at a frame length of 25 ms, with pre-emphasis coefficient of 0.97. For each phoneme, 3-state HMMs with 75 mixtures were used.

Recognition accuracies obtained with the TIMIT and NTIMIT databases at various SNRs were depicted in Figure 3. The accuracy ranges from 67.9% (clean TIMIT) to 34.1% (NTIMIT at 0 dB SNR).

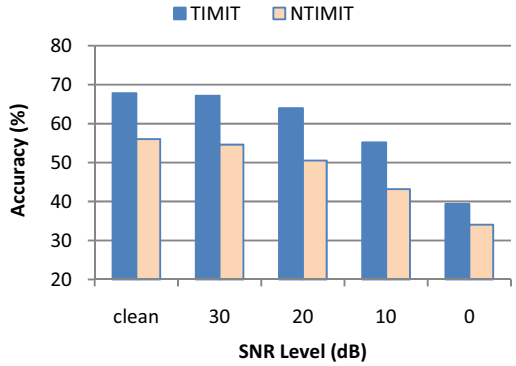

Figure 3: *Accuracies with control MFCC features.*

3.2. Experiment 2: High resolution spectral features

The recognition using high resolution DCTC features only was evaluated in this experiment. 20 DCTCs were computed using 25 ms frames and the Mel-like warping given in Equation (3). This is considered the spectral only case as no DCSC terms were used.

The same configuration of HTK as for Experiment 1 was used, except 25 mixtures were used to model each Gaussian, as errors were obtained using 32 or more mixtures.

Figure 4 shows that phoneme recognition accuracy for the TIMIT and NTIMIT databases at various SNRs, ranging from 55.5% (clean TIMIT) to 20.8% (NTIMIT at 0dB SNR).

These results are considerably lower than the ones obtained in the control experiment, but still about 3% to 6% higher than results obtained with static only MFCC features obtained in another (not reported) experiment.

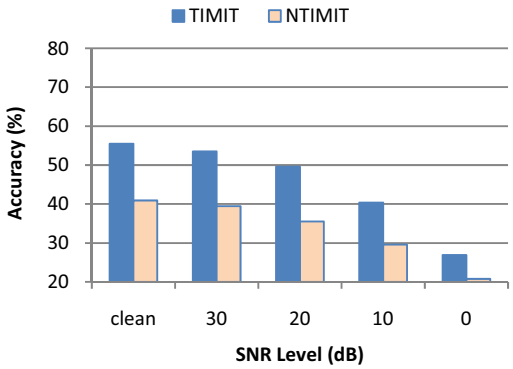

Figure 4: *Accuracies with spectral features only.*

3.3. Experiment 3: Modulation spectrum features

The intent of this experiment was to determine accuracy possible using DCSC features computed to emphasize the modulation spectrum. The DCSC features were computed for 6 DCSCs with 10 DCTC terms for each (60 total features) using Equation (4). Based on several pilot tests, and also with the idea of using low frequency resolution for this evaluation, the spectra were computed with 6 ms frames, spaced 2 ms apart, and the DCTCs were computed using the same frequency warping as used for the spectral only features.

In order for DCSCs to represent modulation frequencies as low as approximately 2 Hz, a 500 ms block length was used for the DCSC calculations (250 frames). The DCSC calculations were performed using basis vectors as shown in Figure 2. These DCSC calculations were repeated with a block spacing of 4 frames (8 ms). For these experiments, the HMMs were configured with 32 mixtures per state.

As shown in Figure 5, phoneme recognition accuracies for the TIMIT and NTIMIT databases at various SNRs range from 72.1% (clean TIMIT) to 37.7% (NTIMIT at 0dB SNR). Compared to spectral features only, the accuracies of the modulation spectrum features increased about 17% for all the cases.

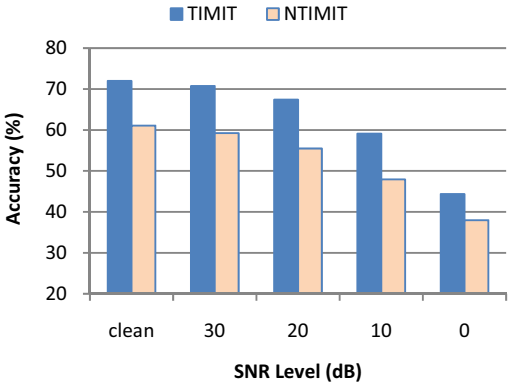

Figure 5: *Accuracies with modulation features.*

3.4. Experiment 4: Combined spectral and modulation features

This experiment was conducted in order to determine if there was an advantage to combine the high resolution spectral features with the modulation spectrum features. Therefore, the 20 high resolution spectral features as used for Experiment 2 were combined with the 60 modulation spectrum features used for Experiment 3, resulting in a total of 80 features. 75 mixtures were used for each state of the HMM models.

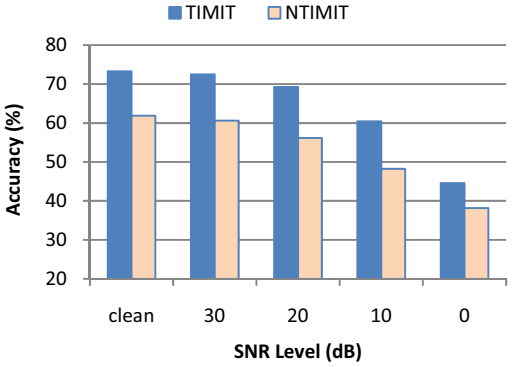

Figure 6: *Accuracies with combined spectral and modulation feature.*

Phoneme recognition accuracies, depicted in Figure 6 for TIMIT and NTIMIT at various SNRs, ranges from 73.3% (clean TIMIT) to 38.2% (NTIMIT at 0dB SNR). The combination of spectral and modulation features leads to only a small improvement in the accuracy compared with that obtained using only the modulation spectrum features.

3.5. Experiment 5: Integrated spectral and modulation features

The intent of this experiment was to examine ASR accuracy using DCTC/DCSC features designed to capture both spectral and modulation information. For this case 13 DCTCs were computed using 8 ms frames spaced 2 ms apart. 6 DCSC terms were computed for each DCTC, using the basis vectors as illustrated in Figure 2, and again using 250 frames per block (500 ms). Thus 78 features were computed per block, with the block spacing 4 frames or 8 ms.

Phoneme recognition accuracies for the TIMIT and NTIMIT database are depicted in Figure 7. The highest accuracy of 73.85% was obtained for the case of clean TIMIT.

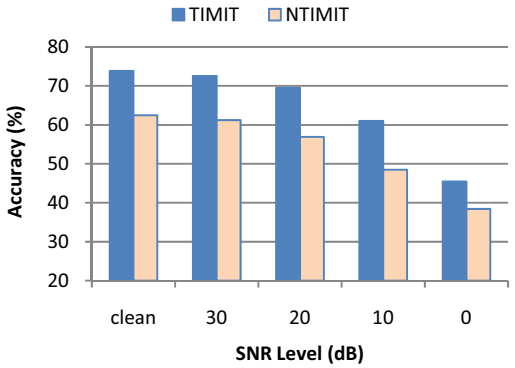

Figure 7: *Accuracies with integrated features.*

4. Literature based Comparison

Table 1 summarizes the best results (clean speech) obtained in this paper with some other recently reported results.

Table 1: *Results reported in literature*.

Feature	Recognizer	Accuracy	Study
TIMIT Results			
PLP	ANN/HMM	71.50%	Ketabdar et al. (2008) [9]
MFCC	GMM	70%	Sha and Saul (2006) [10]
DCT/DCS	HMM	73.85%	this study
NTIMIT Results			
MFCC	HMM	58.79% *	Morales et al. (2007) [11]
DCT/DCS	HMM	62.50%	this study

* A 51-phoneme set and full NTIMIT were used.

5. Discussion

In terms of noise robustness, the modulation features are far more noise robust than spectral features. For example, the spectral only features result in 20.8% accuracy for 0 dB SNR telephone speech versus 38% for the same speech using modulation features. For clean speech, the spectral only features result in reasonable ASR phonetic accuracy (55.5%), but not nearly as high accuracy as that obtained with modulation only features (72.1%). For clean speech and for all noise levels examined, the combination of spectral features and modulation features is marginally better than using modulation features alone. The integrated approach of representing spectral information with medium spectral resolution and detailed temporal information was the best overall, and (slightly) preferred over a combination of high resolution spectral information with modulation features.

6. Conclusions

Most typically, the modulation spectrum information is computed by first dividing the spectrum into individual frequency bands, and then extracting temporal information with a second bandpass filter bank operating over each band. In the approach presented in this paper, the modulation information is obtained by extracting the temporal trajectories of integrated frequency domain features. That is, we represent the temporal trajectories of DCTCs, each of which is obtained from a linear combination of the entire log magnitude spectrum of a frame of speech data. The resultant spectral/temporal features result in higher ASR accuracy than that possible with even very high dimensional more conventional features.

Phonetic recognition results using both the TIMIT and NTIMIT databases compare favorably with any results reported in the literature using these databases. Another contrast between the methods reported here is the relatively poor frequency resolution (8 ms frame length) used for the highest ASR accuracy.

7. References

[1] Hermansky, H. and Morgan, N., "RASTA Processing of Speech," IEEE Trans. Speech and Audio Processing, 2(4), pp.578-589, 1994.

[2] Kanedera, N., Arai, T., Hermansky, H., and Pavel, M., "On the Importance of Various Modulation Frequencies for Speech Recognition," Proc. EUROSPEECH 1997, pp.1079-1082, 1997.

[3] Kanedera, N., and Hermansky, H., "On Properties of Modulation Spectrum for Robust Automatic Speech Recognition," Proc. ICASSP 1998, pp.613-616, 1998.

[4] Kingsbury, B. E. D., Morgan, N., and Greenberg, S., "Robust Speech Recognition using the Modulation Spectrogram," Speech Comm. 25, pp.117-132, 1998.

[5] Valente, F., and Hermansky, H., "Hierarchical and Parallel Processing of Modulation Spectrum for ASR Applications," Proc. ICASSP 2008, pp.4164-4168, 2008.

[6] Zahorian, A. S, Silsbee, P., and Wang, X., "Phone Classification with Segmental Features and a Binary-Pairpartitioned Neural Network Classifier," Proc. ICASSP 1997, pp.1011-1014, 1997.

[7] Karjanadecha, M., and Zahorian, A. S., "Signal Modeling for High-Performance Isolated Word Recognition," IEEE Trans. on Speech and Audio Processing, 9(6), pp.647-654, 2001.

[8] Lee, K. F., and Hon, H. W., "Speaker-independent Phone Recognition using Hidden Markov Models," IEEE Trans. on Acoustic, Speech and Audio Processing, 37(11), pp.1641-1648, 1989.

[9] Ketabdar, H., and Bourlard, H., "Hierarchical Integration of Phonetic and Lexical Knowledge in Phone Posterior Estimation," Proc. ICASSP 2008, pp. 4065-4068, 2008.

[10] Sha, F., and Saul, L.K., "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," Proc. ICASSP 2006, Vol.1, 2006.

[11] Morales, M., Toledano, D. T., Hansen, J. H. L., and Garrido, J., "Multivariate Cepstral Feature Compensation on Band-limited Data for Robust Speech Recognition," Proc. NODALIDA-2007, pp.144-151, 2007.