SUBBAND FEATURE EXTRACTION USING LAPPED ORTHOGONAL TRANSFORM FOR SPEECH RECOGNITION

Z. Tufekci, J.N. Gowdy

Department of Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA
ztufekc@ces.clemson.edu, jgowdy@ces.clemson.edu

ABSTRACT

It is well known that dividing speech into frequency subbands can improve the performance of a speech recognizer. This is especially true for the case of speech corrupted with noise. Subband (SUB) features are typically extracted by dividing the frequency band into subbands by using non-overlapping rectangular windows and then processing each subband's spectrum separately. However, multiplying a signal by a rectangular window creates discontinuities which produce large amplitude frequency coefficients at high frequencies that degrade the performance of the speech recognizer. In this paper we propose the Lapped Subband (LAP) features which are calculated by applying the Discrete Orthogonal Lapped Transform (DOLT) to the mel-scaled, log-filterbank energies of a speech frame. Performance of the LAP features was evaluated on a phoneme recognition task and compared with the performance of SUB features and MFCC features. Experimental results have shown that the proposed LAP features outperform SUB features and Mel Frequency Cepstral Coefficients (MFCC) features under white noise, band-limited white noise and no noise conditions.

1. INTRODUCTION

Conventional feature extraction methods use the entire frequency band to extract speech features for speech recognition. However, as pointed out by Fletcher [1] (and reviewed by Allen in [2]), the Human Speech Recognition (HSR) system works with partial recognition information across frequency, probably in the form of speech features that are local in frequency. Fletcher's work [1] led to the subband-based speech recognizer [3, 4]. Hermansky et al. [4] and Bourlard et al. [3] also proposed subband-based speech recognition systems. They simply divided the frequency band into subbands, extracted features for each subband and then calculated scores for each subband. Finally, they combined each subband's recognition score by using merging techniques. There are three main motivations for the subband-based recognizer:

1. Some subbands of the speech spectrum may be inherently more relevant than others to the task of speech recognition. Therefore, the contribution of each subband to the overall recognition decision can be weighted depending on the information that each subband conveys.

2. Transitions between more stationary segments of speech do not necessarily occur at the same time across the different frequency bands. The subband-based approach may have the potential of relaxing the synchrony inherent in current HMM systems. A frame of speech may contain information of two adjacent phonemes. If one of these phonemes is voiced and the other is unvoiced, then the low-frequency spectrum is dominated by voiced phoneme information, and the high-frequency spectrum is dominated by the unvoiced phoneme information. In traditional feature extraction methods which are based on extracting speech features using the full frequency band, we inherently assume that a speech frame conveys information on only one phoneme at a time. However, this asynchrony can be taken into account by dividing the frequency band into subbands and processing each subband separately.

3. Since the full spectrum of the speech signal is used to calculate feature vectors for full-band recognizers, corruption of a frequency band of speech by noise affects all coefficients. However, corruption of a frequency band affects only a few coefficients if we use a subband based recognizer. Therefore, we can decrease the effect of noise on the performance of the subband-based recognizer by down weighting the contributions from the corrupted subbands.

After pioneering works of Hermansky et al. [4] and Bourlard et al. [3] on the subband-based recognizer, its advantages over full-band based recognizers have been studied in [3–10]. Asynchrony between frequency bands was studied in [3, 5, 8], and it has been shown that accommodating asynchrony between frequency bands improves performance. It also has been shown [4, 7] that subband-based recognizers are robust to band-limited noise but not good for white noise. In summary, subband-based recognizers have three main advantages over full-band based recognizers: (1) they are robust to band-limited noise, (2) they have the ability to incorporate asynchrony between subbands, and (3) they have the ability to weight the contribution of each subband to the total recognition score. The first step in extracting speech features for subband based recognizers is to divide the full speech spectrum into subbands. This is typically done by dividing the full frequency band into subbands using rectangular windows. However, the use of rectangular windows creates large variations of some of the subband coefficients. In this paper we proposed to use the DOLT which uses smooth windows to overcome the problems caused by using rectangular windows.

This work was partially supported by Izmir Institute of Technology

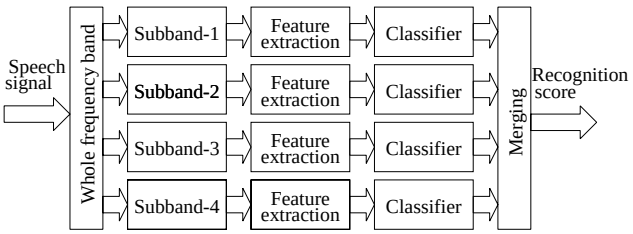

Figure 1: Sub-band based speech recognizer

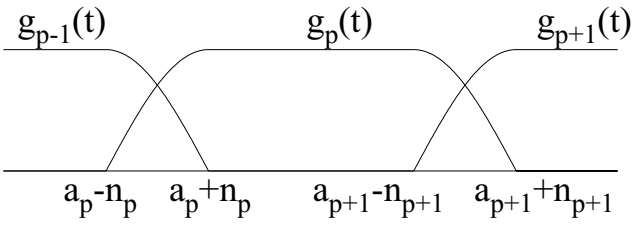

Figure 2: Division of the signal using smooth overlapping windows

2. SUBBAND-BASED RECOGNIZER AND FEATURE EXTRACTION USING THE LAPPED ORTHOGONAL TRANSFORM

In this section the subband-based recognizer [3, 4, 9, 10] is explained first. Then feature extraction using the Lapped Orthogonal Transform [11] is presented. Figure 1 depicts the subband-based speech recognition system. As seen in Figure 1, the frequency spectrum of the speech signal is divided into subbands that may overlap. Then speech features from each subband are extracted. Next, a likelihood for each subband is calculated. Finally, recognition scores from each subband are merged to give the total score. The most critical part of the subband based recognizer is the merging algorithm. The merging algorithm weights the partial recognition scores of subbands based on information conveyed in subbands and/or the signal-noise ratio of each subband. In this paper all subband scores are weighted equally. Since the purpose of this paper is to investigate the effect of using smooth windows instead of rectangular windows to divide the frequency band into subbands, the subband model was kept simple. For example, we did not weight the score of the subbands, and asynchrony between subbands is not incorporated into recognizer.

The Lapped Orthogonal Transform is an alternative to the Block Transform. The Block Transform of a signal is calculated by first dividing the signal into blocks by using nonoverlapping rectangular windows and then transforming each block. The Block Transform has an important disadvantage because of the large variations in the resulting frequency domain coefficients due to the use of rectangular windows. The Lapped Orthogonal Transform uses smooth overlapping windows instead of nonoverlapping rectangular window to divide the signal into blocks. It has been shown [11] that the basis functions of the Lapped Orthogonal Transform can be obtained from the basis function of an orthogonal transform as follows.

Let $\{e_k(t)\}_{k\in\mathbb{N}}$ be a basis of $L^2[0,1]$ (space of the square integrable functions of the interval $[0,1]$). Let g_p be the window of the p'th block with support $[a_p - n_p, a_{p+1} + n_{p+1}]$ (see Figure 2). Also, define $l_p = a_{p+1} - a_p$. Let

$$g_p^2(t) + g_{p+1}^2(t) = 1 \text{ for } t \in [a_{p+1} - n_{p+1}, a_{p+1} + n_{p+1}]. \quad (1)$$

Define

$$\tilde{e}_k(t) = \begin{cases} e_k(t) & \text{if } t \in [0,1], \\ e_k(-t) & \text{if } t \in [-1,0), \\ -e_k(2-t) & \text{if } t \in (1,2], \\ -e_k(2+t) & \text{if } t \in [-2,-1), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Then the family

$$\left\{ g_{p,k}(t) = g_p(t) \frac{1}{\sqrt{l_p}} \tilde{e}_k\left(\frac{t-a_p}{l_p}\right) \right\}_{k\in\mathbb{N}, p\in\mathbb{Z}} \quad (3)$$

is an orthonormal basis of $L^2(\mathbb{R})$ (space of the square integrable functions). $\mathbb{Z}$ denotes integers, and $\mathbb{N}$ denotes positive integers including 0.

The next step is to choose a basis $\{e_k(t)\}_{k\in N}$. The basis functions of the Orthogonal Lapped Transform will be discontinuous, which is not desirable, if the $e_k(t)'$ are not properly chosen. The cosine-IV basis functions are good choices as the $e_k(t)$ in that the resulting basis functions $g_{p,k}(t)$ are continuous. The cosine-IV basis of $L^2[0,1]$ is given below.

$$e_k(t) = \left\{ \sqrt{2} cos[(k+1/2)\pi t] \right\}_{k\in\mathbb{N}} \quad (4)$$

Any window that satisfies (1) can be used for the overlapping windows. Since our signal is discrete, we need the discrete version of the Lapped Orthogonal Transform. The discrete version can be obtained [11] by replacing the orthogonal basis of $L^2[0,1]$ with a discrete basis of $\mathbb{C}^n$ (the set of n-tuples of complex numbers) and uniformly sampling the overlapping windows $g_p(t)$. Let $\{a_p\}_{p\in\mathbb{Z}}$ be a sequence of half integers, $a_p + 1/2 \in \mathbb{Z}$ with $\lim_{p\to-\infty} a_p = -\infty$ and $\lim_{p\to+\infty} a_p = +\infty$. Let $\{e_{k,l}[n]\}_{0\le k<l}$ be an orthogonal basis of signals defined for $0 \le n < l,\ l, n \in \mathbb{Z}$, and let $g_p[n] = g_p(t)_{t=n}$. Define

$$\tilde{e}_k[n] = \begin{cases} e_{l,k}[n] & \text{if } n \in [0, l-1], \\ e_{l,k}[-1-n] & \text{if } n \in [-l, -1], \\ -e_{l,k}[2l-1-n] & \text{if } n \in [l, 2l-1], \\ -e_{l,k}[2l+n] & \text{if } n \in [-2l, -l-1], \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Then the family

$$\left\{ g_{p,k}[n] = g_p[n] \tilde{e}_{k,l_p}[n - a_p] \right\}_{0\le k<l_p, p\in\mathbb{Z}} \quad (6)$$

is a lapped orthonormal basis of $l^2(\mathbb{Z})$ (the space of square summable sequences). In this paper we used Discrete Cosine-IV basis vectors for $e_{l,k}[n]$. The discrete, translated versions of sin and cos were used as overlapping parts of windows. If we use Discrete Cosine-IV basis vectors as $e_{l,k}[n]$, the discrete lapped orthogonal basis vectors will be as follows.

$$\left\{ g_{p,k}[n] = g_p[n] \sqrt{\frac{2}{l_p}} cos[\pi(k+1/2)\frac{n-a_p}{l_p}] \right\}_{0\le k<l_p, p\in\mathbb{Z}} \quad (7)$$

When the signal is of finite length as in our case, the left side of the left-most window and the right side of the right-most window still will have abrupt transitions which will cause large variations in the high frequency coefficients (which represent the left-most and the right-most parts of the signal). The typical practice to reduce these artifacts is to use the Cosine-IV basis for the left most window and use the Cosine-I basis for the right most window [11]. When we use the Discrete Cosine-I basis, the discrete lapped orthogonal basis vectors are as follows.

$$\left\{ g_{p,k}[n] = g_p[n]\sqrt{\frac{2}{l_p}}\lambda_k cos[\pi k \frac{n-a_p}{l_p}] \right\}_{0 \le k < l_p, p \in \mathbb{Z}} \quad (8)$$

where $\lambda_k = 1$ for $0 < k < l_p$ end $\lambda_k = \frac{1}{\sqrt{2}}$ for $k = 0$.

3. EXPERIMENTAL SETUP AND TASK

3.1. Task and Database

We used the TIMIT [12] database to evaluate and compare the performance of the proposed features with MFCC subband features on a phoneme recognition task. 22 phone labels of 61 quasiphonemic labels defined in the TIMIT database were merged into the remaining 39 as in [13]. 39 separate categories are {uw ux}, {uh}, {ah,ax,ax-h}, {aa,ao}, {ae}, {eh}, {ih, ix}, {ey}, {iy}, {y}, {ay}, {ow}, {aw}, {oy}, {er, axr}, {r}, {l,el}, {w}, {m, em}, {n, en, nx}, {ng, eng}, {dx}, {v}, {th}, {dh}, {hh, hv}, {z}, {s}, {bcl, dcl, kcl, pcl, tcl, epi, ps, q, pau}. The confusions within the same categories are not counted in calculating classification accuracy. The sentences which are common to all speakers (labeled "sa" in the TIMIT database) are not used to avoid possible bias towards certain phones. We used complete training and test sets defined in the TIMIT database. There are 168 speakers and 1,344 sentences in the test set, and 442 speakers and 3,536 sentences in the training set. Phones less than two frames in duration were not used. Speakers of the training set and test set are disjoint.

Three-state, left-to-right no-skip, content independent HMM models were constructed for each phoneme category. The output probability distribution of each state was modeled by a mixture of five multivariate Gaussian density functions with diagonal covariance matrix. HTK [14] software was used for training and recognition.

3.2. Feature Extraction

The speech signal sampled at 16 kHz is analyzed with a 32 ms Hamming window every 10 ms. The FFT of each frame is taken to calculate the power spectrum of the signal. For the computation of mel-scaled, log-filterbank energies, 32 triangular mel-scaled bandpass filters were designed [14]. The definitions of feature vectors are given below.

1. MFCC: MFCC [15] are computed by taking the DCT of the mel-scaled, log-filterbank energies. The first sixteen (1-16) of the MFCC are used.
2. SUB-i Features (subband features): Mel-scaled, log-filterbank energies are divided into subbands using non-overlapping rectangular windows. Then each subband's MFCC are calculated. SUB-i denotes the feature vector, calculated by dividing the frequency band into i subbands using nonoverlapping rectangular windows and then determining the MFCC of each subband. The vector size is 16 for all SUB-i features. We attempted to evenly divide the frequency band into subbands. The division of the frequency band into subbands resulted in the following subbands: SUB-2(16, 16), SUB-3(12, 10, 10), SUB-4(8, 8, 8, 8), SUB-5(8, 6, 6, 6, 6), SUB-6(6, 6, 6, 6, 4, 4), SUB-7(6, 6, 4, 4, 4, 4, 4), SUB-8(4, 4, 4, 4, 4, 4, 4, 4). (The numbers inside the parenthesis show the division of the frequency band). For example, the notation SUB-2(16, 16) means that the first band includes 16 log-filterbank energies and the second band includes 16 log-filterbank energies.
3. LAP-i Features: The LAP-i features are subband features, calculated by taking the mel-scaled, log-filterbank energies. LAP-i denotes the feature vector obtained by dividing the speech spectrum into i subbands using smooth overlapping windows and then taking Discrete Cosine-IV or Discrete Cosine-I Transform of the subbands. Figure 3 shows the windows used to divide the mel-scaled, log-filterbank energies into subbands using smooth overlapping windows. The vector size is 16 for all LAP-i features. Since the means of the basis vectors of DOLT are non zero, the mean of each subband is removed before taking the DOLT of mel-scaled, log-filterbank energies. Otherwise, subband coefficients will have extra variation due to the DC value of each subband that may degrade the performance of the recognizer.

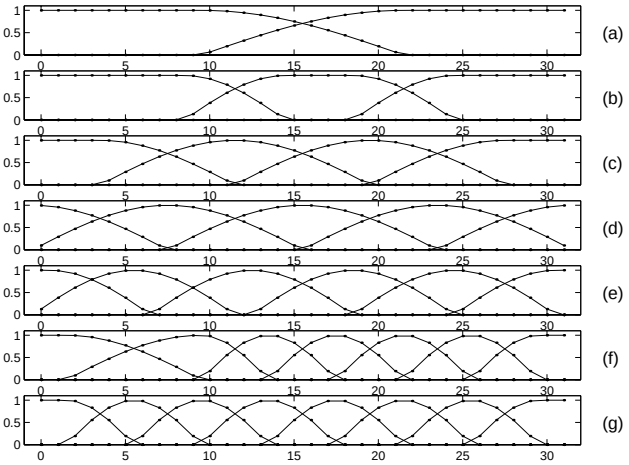

Figure 3: Overlapping windows for LAP feature extraction. (a) to (g) show overlapping windows for LAP-2 to LAP-7. The "." shows the sampling points of the windows.

All feature vectors also include delta coefficients and delta energy to represent dynamic characteristics. Lp-white noise denotes low-pass filtered, white noise with a cut-off frequency of 1.2 kHz.

Table 1: Phoneme recognition rates for MFCC and SUB features for clean and noisy speech

	MFCC	SUB-2	SUB-3	SUB-4	SUB-5	SUB-6	SUB-7	SUB-8
Clean	59.68	60.17	59.50	58.11	57.73	56.20	55.63	54.00
20-db white noise	47.33	50.81	49.97	47.55	47.38	45.24	44.06	42.23
10-db white noise	21.27	23.42	27.01	24.57	24.73	24.28	23.00	20.14
20-db lp white noise	54.62	54.95	55.49	54.66	53.69	51.06	50.49	49.15
10-db lp white noise	45.78	48.68	48.04	44.44	41.83	41.46	40.46	39.08
5-db lp white noise	37.10	40.45	39.65	28.89	27.43	31.45	30.18	29.31

Table 2: Phoneme recognition rates for LAP features for clean and noisy speech

	LAP-2	LAP-3	LAP-4	LAP-5	LAP-6	LAP-7	LAP-8
Clean	61.27	60.52	60.56	60.84	60.57	60.10	59.62
20-db white noise	51.13	50.90	51.27	52.67	51.68	50.82	50.70
10-db white noise	23.57	27.45	26.95	28.91	28.91	29.49	28.36
20-db lp white noise	56.04	57.21	56.73	56.45	56.58	54.15	55.47
10-db lp white noise	49.39	51.42	48.23	48.70	47.01	48.14	47.68
5-db lp white noise	41.53	40.69	36.66	37.77	33.55	38.71	34.35

4. EXPERIMENTAL RESULTS

Table 1 shows the recognition rates for MFCC and SUB features for different noise conditions, and Table 2 shows the recognition rates for LAP features for different noise conditions. As seen from the Table 1 and Table 2 the proposed LAP features yielded better results than the SUB features for a given number of subbands for all noise conditions. The other observation is that recognition rates for SUB features, unlike those for LAP features, decreased drastically as we increased the number of subbands. The number of subbands is important for dealing with noisy speech, since in general, we have better control on decreasing the effect of noise if we divide the frequency band into more slots. From Table 1 and Table 2, we can conclude that for clean speech LAP-3 features are best. However, LAP-7 or LAP-8 features are more effective for noisy speech.

5. CONCLUSION

A new feature extraction method has been proposed for a subband-based recognizer that uses the Discrete Orthogonal Lapped Transform. This method uses smooth overlapping discrete windows to divide the frequency band into subbands. Experimental results have shown that the proposed features consistently yield better results than traditional subband features for a given number of subbands, under all noise conditions investigated. The number of subbands is important for dealing with noisy speech. More subbands allow more control on the effect of noise on the recognition rate. One important advantage of the proposed features over the traditional subband features is that the recognition rate does not decrease significantly as the number of subbands is increased.

6. REFERENCES

[1] H. Fletcher, *Speech and Hearing in Communication*. 1953.

[2] J. B. Allen, "How Do Humans Process and Recognize Speech," *IEEE Transactions on Speech, and Audio Processing*, vol. 2, no. 4, 1994.

[3] H. Bourlard and S. Dupont, "A new ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," in *Proceedings of ICSLP*, 1996.

[4] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on Partially Corrupted Speech," in *Proceedings of ICSLP*, 1996.

[5] N. Mirghafori and N. Morgan, "Transmission and Transition in Multi-band ASR," in *Proceedings of ICASSP*, 1998.

[6] C. Cerisara, J. P. Haton, J. F. Mari, and D. Fohr, "A Recombination Model for Multi-band Speech Recognition," in *Proceedings of ICASSP*, 1998.

[7] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band Speech Recognition in Noisy Environments," in *Proceedings of ICASSP*, 1998.

[8] M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland, and M. A. Fawley, "Modelling Asynchrony in Speech Using Elementary Single-Signal Decomposition," in *Proceedings of ICASSP*, 1997.

[9] H. Bourlard and S. Dupont, "Subband-Based Speech Recognition," in *Proceedings of ICASSP*, 1997.

[10] S. Tiberwala and H. Hermansky, "Sub-band Based Recognition of Noisy Speech," in *Proceedings of ICASSP*, 1997.

[11] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[12] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proceedings of the DARPA Speech Recognition Workshop*, 1986.

[13] R. Chengalvarayan and L. Deng, "Use of Generalized Dynamic Feature Parameters for Speech Recognition," *IEEE Transactions on Speech, and Audio Processing*, vol. 5, May 1997.

[14] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. version 2.1: Entropic Cambridge Research Laboratory Ltd., 1997.

[15] S. B. Davis and P. Mermelstein, "Comparision of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, 1980.