

An Utterance Recognition Technique for Keyword Spotting by Fusion of Bark Energy and MFCC Features*

K. GOPALAN, TAO CHU, and XIAOFENG MIAO
 Department of Electrical and Computer Engineering
 Purdue University Calumet
 Hammond, IN 46323
 U.S.A.

gopalan@calumet.purdue.edu <http://www.calumet.purdue.edu/ece/Gopalan-new.html>

Abstract: This paper describes the preliminary results of a keyword spotting system using a fusion of spectral and cepstral features. Spectral energy in 16 bands of frequencies on Bark scale and 16 mel-scale warped cepstral coefficients are used independently and in combination with appropriate weights for recognizing word utterances. Results of matching features using Euclidean and cosine distances in a dynamic time warping (DTW) process demonstrate that cosine distance works better for Bark energy features while weighted Euclidean distance brings out the closeness of utterances in the cepstral domain. In both cases, performance of DTW shows an accuracy of better than 81 percent for different speakers while fusion of the two feature sets raises the score to over 86 per cent, both based on a small subset of utterances from the Call Home database.

Key-Words: Speech recognition, Bark energy, Mel cepstrum, Feature fusion, Dynamic time warping.

1 Introduction

The goal of a keyword spotting system is to recognize the presence of a small set of a pre-determined words in a continuous stream of speech. The process involves recognizing selected keywords in speech utterances containing extraneous (out of vocabulary) speech and noise. For a successful system, it is important to obtain the highest possible recognition score without incurring incorrect detection of non-keywords or false negatives. A keyword spotting system has a variety of applications in human-machine interaction, some of which include voice-activated telephone assistance, news and video mail retrieval, and monitoring of criminal suspects and prison inmates from telephone calls.

Recognizing one or more words in an unconstrained continuous stream of speech entails tasks such as word boundary detection with coarticulation. In addition, video- and news-indexing and monitoring of telephone conversations generally involve out-of-vocabulary keywords such as acronyms, names and foreign words along with homonyms, which

render the recognition more difficult than commonly used words.

Speech recognition techniques, in general, follow a template matching of features with time normalization by dynamic time warping [1, 2]. Features used for creating templates are typically derived from spectral or log spectral representation of each frame of speech. Parameters from a linear prediction model and mel-cepstral coefficients have been commonly used for forming templates. Mel-frequency cepstral coefficients (MFCCs), in particular, have been proven to be an excellent feature set for isolated word speech recognition [3]. In another development, statistical parametrization of keyword utterances using linear predictive and cepstral coefficients was employed in a hidden Markov model (HMM) to achieve close to 90 percent recognition accuracy. More recently, HMM-based approach has been widely used with phonemic garbage models for non-keyword intervals [4, 5]. Due to the large amount of training data required for efficient HMM representation of keyword, however, dynamic time warping, in spite of its large

* This work is supported by a grant award from the Air Force Research Laboratory, Rome, NY, U.S.A. with the award number FA8750-08-2-0103.

computational requirement, is still considered a viable alternative [6].

It is evident that regardless of the matching technique employed, keyword recognition scores depend on (a) the efficacy of the parameters representing utterances so that they discriminate between different words while appearing close for the same words regardless of speaker, (b) the measure of dissimilarity that effectively accentuates the difference between two different words in the feature domain, and (c) the time aligning process that takes into account the difference in durations between two utterances. In this paper, we present the preliminary results of combining two sets of perception based features, namely, Bark scale based energy parameters and MFCCs. The fusion of the two sets of parameters with appropriate weights for each set is used in a dynamic time warping process for template matching.

In the following sections we discuss the features employed from spectral and cepstral representation of speech. We then present and discuss the results of using a dynamic time warping process for template matching of features from reference keywords and unknown utterances. We conclude the paper with the results of the effect of fusion of the two sets of features with appropriate measures of dissimilarity.

2 Bark Energy and Cepstral Features of Speech

Critical bands of frequencies, as are well known, are used as a measure to relate human auditory perception and frequency resolution of loudness, pitch, etc. Bark frequency scale, which represents a set of filters with bandwidths equal to critical band of perception, is a class of critical band scales that approximates the spectrum analyzing behavior of the cochlea. The filter bands of the cochlea are represented by the Bark scale with finer perceptual resolution at lower frequencies and increasing in bandwidth at higher frequencies. Spectral energy over the Bark scale is considered more natural in approximating perception in the ear [7]. While the Bark band covers the entire audio range, the edge frequencies start near zero and go up to 4000 Hz in nonoverlapping bands for

utterances sampled at the rate of 8000 per second.

Inasmuch as acoustic characteristics are better represented in the spectral domain, another set of perceptually based features is obtained from mel scale-based cepstral representation [7]. The commonly used mel scale, which is approximately linear up to 1000 Hz and logarithmic thereafter, is based on pitch comparisons and is generally considered to characterize speech segments better than linear frequency scale.

Typically, from 13 to 20 MFCCs starting at around 100 Hz and going up to approximately 7000 Hz for speech sampled at 16000 Hz are used as features for speech or speaker recognition [3]. Bandwidth is constant at around 100 Hz for center frequencies up to 1000 Hz and it increases to approximately 950 Hz at the center frequency of 7000 Hz. Critical band filters with these characteristics usually have a triangular response in the spectral domain with constant peak responses at all center frequencies; filters that are increasing in bandwidth (upper band of filters) may also have decreasing peak responses so that each filter response has the same spectral energy.

3 Database Used and Its Feature Domain Representation

With the ultimate goal of recognizing the presence of a selected word – keyword – from an unconstrained speech, utterances from the CallHome database were used for testing the proposed fusion of the Bark energy (BE) and MFCC features. The database consists of speech utterances from telephone conversations between different speakers in two channels, each sampled at the rate of 8000 per second. The following utterances from the database were selected for feature computation and matching: 1. *conversation* (10 utterances from 7 different speakers), 2. *continue* (1), and 3. *unwind* (1). These words were chosen because of their durations being approximately the same. The goal was to determine a threshold of dissimilarity for recognizing the keyword *conversation* independent of the speaker from the above list of 12 utterances. This threshold must yield as few false positive and false

negative recognition scores as possible for the selected set of utterances. Once the threshold is established for a keyword, the same can be used for spotting its presence among other utterances. To reduce processing time and complexity, utterances of durations shorter or longer by 30 percent of the 'nominal' duration for the keyword utterance *conversation* can be eliminated in preprocessing. Selection of 'reference' keyword utterance duration is based on the frequency of occurrence of the utterance from a selected set in the database.

We used the first 16 Bark bands covering up to 3150 Hz while discarding the spectral band of 3150 Hz – 4000 Hz due to its low energy. Also, the discrete Fourier transform (DFT) resolution is used as the starting frequency in place of zero to avoid dc. Each utterance was divided into frames of 128 samples (16 ms) with 64-sample (8 ms) overlap and the spectral energy in each of the 16 bands was obtained with a 512-point DFT giving a frequency resolution of $\Delta f = 8000/512 = 15.625 \text{ Hz}$. Figs. 1 and 2 show the trajectory of Bark energy for the 3rd band (187.5 Hz to 281.25 Hz) for two utterances of the keyword utterance *conversation*, and the two non-keyword utterances, *continue* and *unwind*.

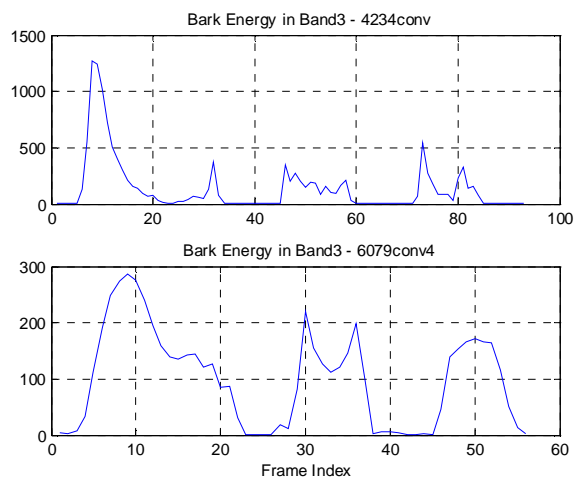


Fig. 1 Bark energy profile for *conversation* by different speakers

As can be seen from the figures, the similarity in the Bark energy patterns across the frames for the same word while differing for different words indicates the value of the Bark energy as a suitable feature element. We

particularly note that while the number of frames is different for the same word – depending on the speaker's rate of speech – the general pattern of energy distribution remains the same. Although mel frequency cepstral coefficients are generally considered to contain more relevant information needed for speech recognition, the Bark energy distribution shows that it can be a supplement, if not an alternative, to other more involved features such as MFCCs or linear prediction based parameters.

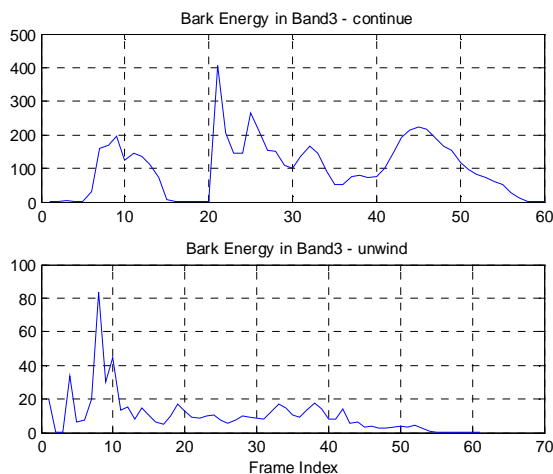


Fig. 2 Bark energy profile for *continue* and *unwind* by different speakers

4 Feature Discrimination Results and Discussion

For the purpose of identifying the threshold for the keyword *conversation*, and to study the effect of the Bark energy features, word utterances in the vicinity of the utterance for *conversation* were located and extracted. Preprocessing by the number of frames based on 128 samples per frame with 64 samples of overlap eliminated all but the three utterances, namely, *conversation* (10), *continue* (1) and *unwind* (1). Bark energy values in the first 16 bands covering the frequency range of 15.625 Hz to 3150 Hz were computed for each frame. The 16 energy values for each frame with variance normalization across all frames were used as a feature for an utterance. A dynamic time warping process using the cosine of two vectors, $A = [a_1 a_2 \dots a_N]^T$, and $B = [b_1 b_2 \dots b_N]^T$ of N elements each was carried out for each pair of normalized features. The

cosine-based distance between A and B is given by

$$d_c = 1 - \cos(\cdot) = 1 - \frac{A^T B}{\sqrt{\sum_{n=1}^N (a_n)^2} \sqrt{\sum_{n=1}^N (b_n)^2}} \quad (1)$$

Dissimilarity between feature vectors corresponding to a pair of utterances is given by the average – based on the shorter utterance of the two – of the total distance in going from the first to the last frame of the shorter utterance. Table I shows the dissimilarity values between pairs of utterances using Bark energy features.

From this table, a value of 0.25 may be chosen for an optimal threshold for recognizing the word *conversation*. At 0.25 for the threshold, the keyword *conversation* is missed in four cases (rows for Conv7 and Conv8) – highlighted in Table 1 – giving a false negative

of 4 out of 9, and four false positive for the two non-keywords. If Conv1 is used as the reference utterance with the threshold of 0.25, no false positive or negative recognition results, as seen from the column under 4234 Conv1. If, for the same threshold, 5254 Conv2 is used as the reference utterance, however, 6079A conv7 is missed but no non-keyword is incorrectly recognized. Overall, for the selected threshold of 0.25, the score is 58/66 \approx 87.8 % with 4/66 \approx 6.1 % for both false negative and false positive. The same features with unweighted and weighted – both linear and sinusoidal – Euclidean distance measures yielded overall recognition scores ranging from approximately 74 % to 77 %.

Table 1
Dissimilarity Measures between Pairs of Utterances using Bark Energy Features and Normalized Cosine Distances

	4234 Conv1	5254 Conv2	4315 Conv3	4431 Conv4	4521 Conv5	4521B Conv6	6079A Conv7	6079B Conv8	6079C Conv9	6079D Conv10	Cont.
Conv2	0.165										
Conv3	0.100	0.160									
Conv4	0.173	0.166	0.181								
Conv5	0.195	0.168	0.203	0.215							
Conv6	0.210	0.206	0.199	0.248	0.130						
Conv7	0.221	0.264	0.200	0.221	0.256	0.296					
Conv8	0.215	0.194	0.195	0.227	0.240	0.253	0.220				
Conv9	0.190	0.207	0.188	0.198	0.243	0.236	0.180	0.123			
Conv10	0.190	0.227	0.225	0.210	0.212	0.212	0.240	0.151	0.182		
continue	0.305	0.268	0.207	0.285	0.228	0.252	0.27	0.252	0.293	0.250	
unwind	0.357	0.284	0.277	0.301	0.252	0.343	0.248	0.292	0.374	0.374	0.290

For comparison, 16 MFCCs were evaluated at the center frequencies of [125 250 375 500 625 750 875 1000 1156.3 1312.5 1500 1718.8 1937.5 2218.8] Hz with bandwidths varying from 250 Hz to 781.2 Hz. Each of the triangular bands of filters has a constant peak response of unity. The MFCC's were evaluated for each frame of an utterance after preemphasizing by a filter of the form $H(z) = 1 - a_k z^{-1}$ (with the coefficient a_k adapted from

frame to frame based on the ratio of frame autocorrelation at unity lag to energy), and

normalized using the variance across all the frames. The variance-normalized MFCC features of all the 12 utterances were compared in the same DTW process as for the Bark energy features using different dissimilarity measures. Table 2 shows the pair wise distances using sinusoidal Euclidean distances measure given by

$$d_{Es} = \sqrt{\sum_{n=1}^N [w(n)a(n) - w(n)b(n)]^2} \quad (2)$$

where the weight vector $w(n)$ is given by

$$w(n) = 1 + \frac{N}{2} \sin\left(\frac{n\pi}{N}\right) \quad (3)$$

With 10.8 as the threshold of dissimilarity in Table 2, the word *conversation* is correctly recognized in all cases, but both the non-keywords also incorrectly fall below the threshold when Conv1 is used as reference; this gives a false positive of 2 and an overall score of 9/11 with the single reference of Conv1. For the

same threshold of 10.8, only one false positive (*unwind*) results, however, with Conv2 used as the reference utterance. With Conv10 as the reference, no false positive or negative score occurs for the same threshold. For the whole set of 66 pairs, the threshold of 10.8 results in the overall recognition score of 54/66 \approx 81.8% with 11 false negative and one false positive.

Table 2
Dissimilarity Measures between Pairs of Utterances using MFCCs and Sinusoidal Euclidean Distances

	4234 Conv1	5254 Conv2	4315 Conv3	4431 Conv4	4521 Conv5	4521B Conv6	6079A Conv7	6079B Conv8	6079C Conv9	6079D Conv10	Cont.
Conv2	8.299										
Conv3	8.519	9.238									
Conv4	8.698	8.562	10.144								
Conv5	10.134	10.242	10.592	9.915							
Conv6	8.992	8.771	7.930	10.526	10.456						
Conv7	9.022	10.758	9.555	9.668	11.868	10.511					
Conv8	8.193	9.240	8.818	9.444	9.403	9.397	10.483				
Conv9	8.337	9.505	9.187	9.726	10.454	8.931	10.660	7.790			
Conv10	8.343	9.930	9.858	9.137	10.198	10.049	10.221	9.622	9.865		
continue	9.401	11.468	9.109	9.939	11.359	8.895	11.683	11.088	11.313	10.957	
unwind	9.730	10.249	8.894	9.894	11.045	10.048	10.972	9.893	9.949	10.916	10.857

The above results show that the Bark energy features are comparable to MFCC features in representing and recognizing utterances. In fact, depending on the dissimilarity measure used, Bark energy features may outperform MFCCs in feature matching. Therefore, a combination of the two feature sets – as an augmented feature vector with 16 energy values and 16 MFCCs – were used in the DTW process for recognizing utterances. The results of the fusion of the two sets of features with a weight of 1.0 for MFCC – due to its greater significance in representing perception – and 0.9 for Bark band energy are shown in Table 3. We note that each feature set is normalized by its variance so that the dissimilarity measures between a pair of augmented feature vectors have approximately the same range.

The overall score with the combination of the two features shows a recognition rate of

57/66 \approx 86.4 per cent. While this score is less than the score obtained using only the Bark band energy features, it is still higher than the most commonly used MFCC features; for a larger word list of more than the 12 words used in the experiment, however, we note that MFCC features may be more dominant in determining the overall recognition accuracy.

5 Conclusion

A method of representing speech utterances using a fusion of Bark band energy and MFCC features has been proposed for a keyword spotting system. The computationally simple Bark band energy feature appears to yield low dissimilarity values between the same utterances independent of speakers. Augmenting the more commonly used MFCC features with the Bark band energy features shows an improvement in the total recognition score for the small size of

word list tested. Further studies using a larger word list with reduced feature size while supplementing the feature vector with time

derivative of the energy and MFCC, etc. are in progress.

Table 3
Overall Dissimilarity Values between Pairs of Utterances using MFCCs (with Sinusoidal Euclidean Distance) and Bark band Energy (with cosine Distance)

	4234 Conv1	5254 Conv2	4315 Conv3	4431 Conv4	4521 Conv5	4521B Conv6	6079A Conv7	6079B Conv8	6079C Conv9	6079D Conv10	Cont.
Conv2	1.046										
Conv3	1.010	1.167									
Conv4	1.089	1.034	1.262								
Conv5	1.298	1.217	1.324	1.252							
Conv6	1.139	1.103	1.059	1.337	1.214						
Conv7	1.183	1.368	1.218	1.225	1.489	1.398					
Conv8	1.084	1.173	1.175	1.217	1.219	1.241	1.354				
Conv9	1.090	1.211	1.157	1.193	1.303	1.165	1.342	0.946			
Conv10	1.110	1.279	1.234	1.155	1.287	1.262	1.392	1.128	1.226		
continue	1.266	1.445	1.242	1.340	1.450	1.225	1.521	1.484	1.503	1.556	
unwind	1.378	1.480	1.294	1.404	1.470	1.394	1.523	1.365	1.421	1.615	1.644

References:

- [1] R. W. Christiansen and C. K. Rushforth, "Detecting and Locating Key Words in Continuous Speech Using Linear Predictive Coding," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. ASSP-25, pp. 361-367, Oct. 1977.
- [2] C. Myers, L. Rabiner, and A. Rosenberg, "An investigation of the use of dynamic time warping for word spotting and connected speech recognition," *Proc. of the IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP '80)*, vol. 5, pp. 173-177, Apr. 1980.
- [3] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol 28, No. 4, pp. 357 - 366, Aug. 1980.
- [4] J.G. Wilpon, L.R. Rabiner, C. -H. Lee, and E.R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 38, Issue 11, pp. 1870-1878, Nov. 1990.
- [5] K. Thambiratnam, and S. Sridharan, "Dynamic Match Phone-Lattice Searches For Very Fast and Accurate Unrestricted Vocabulary Keyword Spotting," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP '05)*, pp. 465-468, Mar. 2005.
- [6] Y. -D. Wu and B. -L. Liu, "Keyword Spotting Method Based on Speech Feature Space Trace Matching," *Proc. IEEE Second Int. Conf. Machine Learning and Cybernetics*, pp. 3188-3192, Nov. 2003.
- [7] X. Huang, A. Acero and H.-W. Hon, *Spoken Language Processing*, Prentice-Hall PTR, Upper Saddle River, NJ, 2001.