

Speech Recognition for Keyword Spotting using a Set of Modulation Based Features – Preliminary Results*

Kaliappan GOPALAN and Tao CHU

Department of Electrical and Computer Engineering
Purdue University Calumet
Hammond, IN 46323

ABSTRACT

We present the preliminary results of applying a set of parameters of the AM-FM model for recognizing word utterances. By acquiring modulation based parameters from the amplitude envelope (AE) and the instantaneous frequency – both obtained by demodulating at four selected center frequencies – a compact feature set is created for each frame of a word utterance. Applying a dynamic time warping of features, a dissimilarity measure between an unknown and one of several reference utterances is obtained to detect the presence of a keyword in a continuous stream of speech. A feature set consisting of the peak frequencies in AE and weighted formants, among others, shows an overall recognition score of 75 percent or higher – depending on the analysis frequencies used – for an extracted set of word utterances. The low false positive and false negative scores suggest the viability of modulation based parameters for building a keyword spotting system.

Key Words: Speech recognition, AM-FM Model, Dynamic Time warping.

1. INTRODUCTION

Keyword recognition is concerned with the detection of a pre-fixed set of words in a continuous stream of speech. The process involves locating the occurrence of selected keywords in speech containing extraneous (out of vocabulary) speech and noise. Prior methods of recognition typically involved template matching of keyword features with time normalization by dynamic time warping [1, 2]. Features used for creating templates are commonly derived from the spectral or log spectral representation of each frame of speech – templates are formed using parameters from linear prediction model and mel frequency cepstral coefficients. Additionally, 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 [3]. More recently, HMM-based approach has been widely used with phonemic

garbage models for non-keyword intervals [4]. Due to the large amount of training data required for efficient HMM representation of keyword, however, dynamic time warping, in spite of its large computational requirement, is still considered a viable alternative [5].

It is clear that regardless of the pattern 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, (b) the measure of dissimilarity that effectively accentuates the difference between two words in the feature domain, and (c) the time aligning process that takes into account the difference in durations between two utterances. In the following sections we present the preliminary results of employing features derived from the modulation model of speech for recognizing a word utterance, independent of speakers, using dynamic time alignment with two measures of dissimilarity.

2. AM-FM MODEL OF SPEECH

Teager and Teager [6] observed that modulation process dominates the production of speech and showed that speech resonances have both frequency modulation and time-varying amplitudes. Other researchers postulated that human auditory system uses transduction of frequency modulation (FM) to amplitude modulation (AM) using the spectral shapes of auditory filters. Based on the results of these works, the many nonlinear and time-varying phenomena during speech production have been modeled successfully by AM-FM models representing each of the resonances or formants. Maragos, et al [7] used the nonlinear Teager energy operators to obtain the instantaneous frequency and the amplitude envelope in the AM-FM model of speech in the vicinity of resonant frequencies or formants.

Because of the nonlinearities in speech and the modulation model is an ill-posed problem, speech signal is bandpass filtered around a resonant frequency $\Omega_c =$

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

$2\pi f_c$ and is modeled as comprising of AM and FM components as given by

$$s(n) = a(n) \cos(\Omega_c n + \Omega_m \int_0^n q(k) dk) \quad (1)$$

Using the algorithm formulated by Kaiser [8], the Teager energy of the bandpass filtered signal $s(n)$ is calculated by the operator $\psi(\cdot)$ given by

$$\psi[s(n)] = s^2(n) - s(n-1)s(n+1) \quad (2)$$

Teager energy, and mel cepstrum based on Teager energy were used in the past to obtain features for isolated word recognition [9, 10].

From the energy operators of a frame of speech signal and its time-shifted versions, the instantaneous frequency (IF) Ω_i and the amplitude envelope (AE) $|a(n)|$ are calculated as

$$\Omega_i(n) \approx \arcsin \sqrt{\frac{\psi[s(n+1)] - \psi[s(n-1)]}{4\psi[s(n)]}} \quad (3)$$

$$|a(n)| \approx \frac{2\psi[s(n)]}{\sqrt{\psi[s(n+1)] - \psi[s(n-1)]}} \quad (4)$$

3. SPEECH FEATURES FROM AM-FM MODEL

Inasmuch as the modulation model brings out the time-varying characteristics of speech production around a formant, features from AE and IF can be used to represent speech. Since AE is a slowly varying component with bandwidth no greater than that of the bandpass filter used (600 Hz to 800 Hz) to obtain $s(n)$, the spectral behavior of AE can form a useful feature. Additionally, if the filtered signal has a resonance in the vicinity of an arbitrarily selected center frequency f_c , the resonant frequency derived from IF and its bandwidth are significant in describing the signal.

With these considerations, first the following parameters were evaluated for use as elements of feature vector for each frame of speech that is obtained at the sampling rate of 8000 Hz. 1. Total energy of AE at each center frequency, 2. Band energy of AE in the 250 Hz – 500 Hz band, 3. the first peak frequency of AE, 4. the unweighted formant estimate, and 5. the weighted formant estimate. Each of these parameters was obtained at four center (approximate resonant) frequencies of 900Hz, 1600Hz, 2500Hz and 3500Hz. Choice of arbitrary frequencies, in general, has been shown to work well for analyzing pitch and voiced/unvoiced decisions without incurring formant calculations [11]. Although very few of the frames of speech in the utterance of a keyword of

interest have these four frequencies as the first four formants, the choice was made for uniformity regardless of the location/absence of the formants. With five parameters at each of the four center frequencies, feature vector for each frame consists of 20 elements. (Other parameters such as the maximum and minimum instantaneous frequencies at each f_c , low frequency energy of AE etc., were considered and discarded as insignificant in characterizing an utterance.) Figure 1 shows the efficacy of the first peak frequency in discriminating different utterances. The pairs of figures in (a) and (b) for the utterances *conversation* and *circumstance*, respectively, indicate the similarity of the feature trajectories for the same words (intra word similarity) while significantly discriminating between the two words (inter word discrimination), all obtained at the analysis frequency of 1600 Hz.

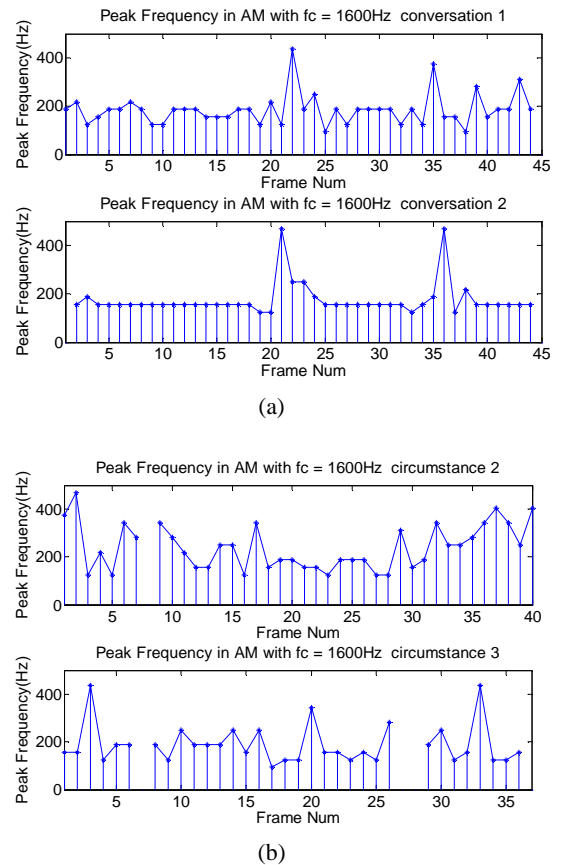


Figure 1. Peak frequencies of AM envelope at the center frequency of 1600 Hz for two utterances each of (a) *conversation*, and (b) *circumstance*

In the second experiment, AE and IF were first obtained at the center frequencies of [550 1600 2600 3500] Hz. From the AE and IF at each center frequency, weighted formant estimates were calculated for all the voiced sections. Next, average of these weighted formants across all the voiced sections, one for each of F1, F2, F3, and F4, was used as the center frequency for

extracting modulation based features. For each utterance, feature vectors comprising of (a) peak frequencies in AE, (b) total Teager energy, (c) AE energy derivative, (d) geometric mean-to-algebraic mean ratio of AE, and (e) weighted formant were obtained at each center frequency. Figure 2 shows the total Teager energy (TTE) profiles for two utterances each of the pair, *conversation*, and (b) *circumstance*. These profiles clearly display the capability of TTE in distinguishing different utterances. Inasmuch as the Teager energy is the basis for AM-FM analysis of speech in Eqs. (3) and (4), this capability of feature discrimination is carried over for the derived modulation parameters such as the amplitude envelope and instantaneous frequency. The derivative of the amplitude envelope (obtained to reveal speech-dependent high frequency variations within the envelope), for example, shows similar intra and inter word profiles as displayed in Figure 3.

Further refinement to the modulation analysis frequencies was carried out in the third experiment by using an average set of the actual formants F1, F2, F3, and F4 for a set of three reference utterances of the selected keyword.

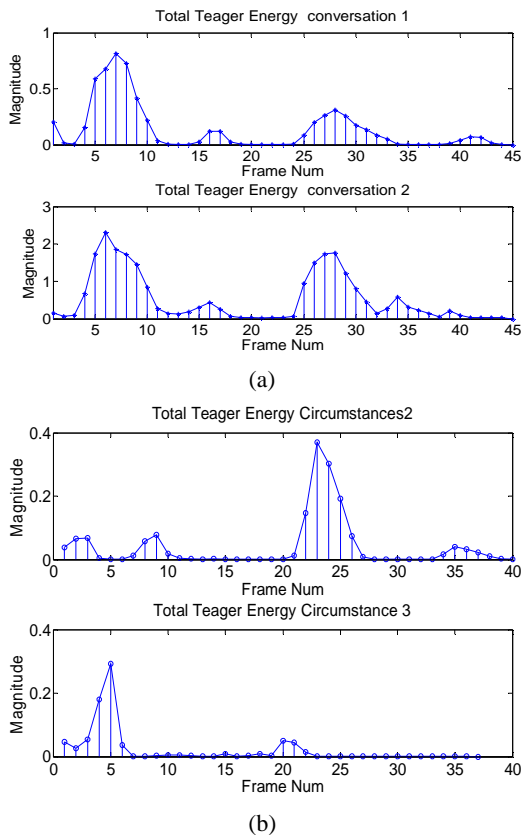


Figure 2. Total Teager energy profiles for two utterances each of (a) *conversation*, and (b) *circumstance*

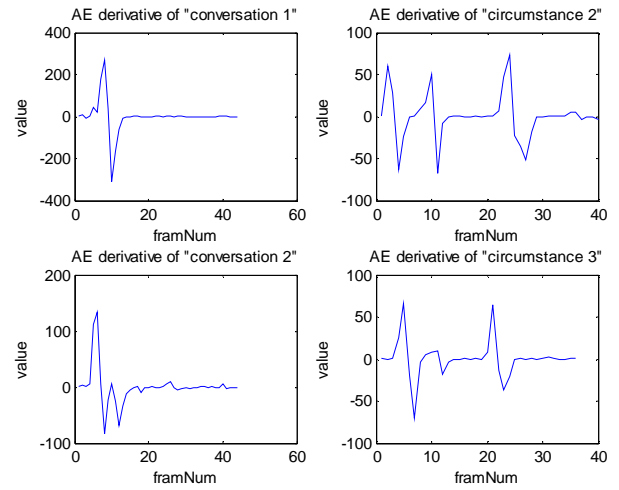


Figure 3. Profiles of amplitude envelope derivative for two utterances each of *conversation*, and *circumstance*

4. RESULTS AND DISCUSSION

The following utterances from the Call Home data base were used for testing the efficacy of the modulation-based feature sets discussed in the preceding section. 1. *conversation* (10 utterances spoken by 7 different female speakers), 2. *circumstance* (3 utterances), 3. *apartment* (1), (4) *continue* (1), and 5. *unwind* (1). These words were chosen because of their durations being approximately the same. The goal was to recognize the keyword *conversation*, independent of the speaker, with one or more utterances of *conversation* used as reference. (Utterances of durations shorter or longer by 50 percent of the average duration for the three reference utterances of *conversation* were eliminated in preprocessing.)

To compare the features and obtain an inter-utterance dissimilarity, a dynamic time warping (DTW) process was used with dissimilarity calculated using the cosine of the angle between two multicomponent vectors. The 20-parameter feature vectors in the first experiment, namely, 1. total energy of AE at each center frequency, 2. band energy of AE in the 250 Hz – 500 Hz band, 3. first peak frequency of AE, 4. unweighted formant estimate, and 5. weighted formant estimate, each obtained at the five center frequencies of 900 Hz, 1600 Hz, 2500 Hz and 3500 Hz, were used in a pair wise comparison in the DTW process with a subset of the 11 test utterances with 7 containing the keyword. Using a DTW distance threshold based on a reference set of three utterances, an overall recognition score of six out of eight resulted with false and false negative of one each. Although this score was reasonable at around 65 percent, it dropped to about 60 per cent when the full list of 16 utterances were used.

In the second experiment, features consisting of the AE energy derivative (AEED), total Teager energy (TTE), the peak frequency of AE (PFAE), weighted

formant (WFMT), and the geometric mean-to-algebraic mean ratio (GAR), were evaluated each at the refined analysis frequencies of [409.9 1678.6 2440.3 2901] Hz. (These resonant (formant) frequencies were obtained from the weighted formants around the arbitrary set of [900 1600 2500 3500] Hz.) A bandwidth of 600 Hz was used at each analysis frequency (except at 409.9 Hz for which it was approximately 400 Hz). The resulting pairwise dissimilarity measures employing cosine of angle between two feature vectors are shown in Table I. From this table, we observe that the features are able to discriminate between an utterance of the keyword *conversation* and the other words that are approximately the same in length.

Comparing the dissimilarity between the keyword utterance Cv1 and each of Cv2, Cv3 and Cv4, a threshold of 0.71 can be used as the largest value between two

utterances of the keyword. At this threshold, with Cv1, Cv2, and Cv3 (each spoken by a different female speaker) as references, an unknown word utterance X may be recognized as the keyword if X has a dissimilarity of below the threshold with at least two out of the three reference utterances. With this simple rule, keyword utterances Cv5, Cv6 and Cv8 are missed while all the non-keyword utterances tested are correctly rejected. (It must be noted that while Cv7 – Cv10 were spoken by the same female speaker, Cv8 corresponded to a mispronunciation of *conversation* as *converstatement*; it was included as a potential keyword to test the feature sets.) If Cv2, Cv3, and Cv4 are used as references, only Cv6 and Cv8 are falsely rejected and no non-keyword is misrecognized. Similar results can be seen with other combinations as references. Comparable dissimilarity values and false positive and negative scores resulted using the Euclidean distance measure.

Table I
Dissimilarity Measures between pairs of Utterances using Cosine of Angle between Feature Vectors

	Cv1	Cv2	Cv3	Cv4	Cv5	Cv6	Cv7	Cv8	Cv9	Cv10	Cont	Unw	Cir1	Cir2	Cir3
Cv2	0.69														
Cv3	0.69	0.71													
Cv4	0.68	0.71	0.68												
Cv5	0.74	0.69	0.74	0.70											
Cv6	0.74	0.68	0.72	0.74	0.77										
Cv7	0.71	0.70	0.70	0.70	0.70	0.73									
Cv8	0.73	0.79	0.72	0.73	0.73	0.70	0.68								
Cv9	0.69	0.78	0.63	0.70	0.71	0.73	0.81	0.64							
Cv10	0.71	0.66	0.68	0.68	0.72	0.69	0.78	0.70	0.70						
Cont	0.78	0.83	0.74	0.73	0.74	0.75	0.70	0.77	0.85	0.75					
Unw	0.80	0.81	0.74	0.68	0.75	0.69	0.75	0.70	0.69	0.74	0.77				
Cir1	0.78	0.70	0.73	0.74	0.67	0.70	0.68	0.62	0.74	0.69	0.80	0.77			
Cir2	0.76	0.79	0.75	0.74	0.72	0.70	0.70	0.74	0.79	0.72	0.76	0.67	0.68		
Cir3	0.76	0.78	0.72	0.74	0.74	0.73	0.74	0.83	0.74	0.71	0.82	0.77	0.69	0.75	
Apt	0.72	0.77	0.71	0.78	0.75	0.72	0.78	0.71	0.74	0.75	0.76	0.74	0.86	0.70	0.77

Utterances Cv1 – Cv10: *conversation*; Con.: *continue*; Unw: *unwind*; Cir1 – Cir3.: *circumstance*; Apt.: *apartment*

In the third experiment, formants in each voiced frame of three reference utterances of the keyword *conversation* (Cv1, Cv2, and Cv3, each by a different female speaker) were first evaluated using linear prediction error, and the averages of the four formants, [F1 F2 F3 F4] = [401.3 1272.2 2219.6 3066.2] Hz, were used as the modulation analysis frequencies with bandwidths of [300 400 500 600] Hz. For utterance comparison, the same features as in the previous experiment, namely AEED, TTE, PFAE, WFMT and GAR, were applied in a DTW process with each pair of utterances. Resulting Euclidean distances between each pair of utterances in the feature domain are given in Table II. In

this case, employing Cv1, Cv2, and Cv3 as references with a threshold of 3.3 gives one false negative (Cv5) and two false positive (Cir1 and Cir3) scores, with the overall recognition score of 9/12 or 75 per cent. Dissimilarities using the cosine measure gave a score of 5 for false negative and one for false positive for the small database used. Although the scores are no better than the ones from the second experiment, employing average formants of reference utterances as analysis frequencies is, in general, physically more meaningful than using arbitrary frequencies; hence, it is expected that higher recognition scores may result for a larger data set.

Table II
Dissimilarity Measures between pairs of Utterances using Euclidean Distance

	Cv1	Cv2	Cv3	Cv4	Cv5	Cv6	Cv7	Cv8	Cv9	Cv10	Cont	Unw	Cir1	Cir2	Cir3
Cv2	3.33														
Cv3	3.13	3.13													
Cv4	3.17	3.14	3.09												
Cv5	3.41	3.22	3.45	3.12											
Cv6	3.37	3.24	3.17	3.36	3.17										
Cv7	3.22	3.41	3.12	3.13	3.11	3.20									
Cv8	3.27	3.49	3.11	3.23	3.23	3.30	3.35								
Cv9	3.29	3.47	2.92	3.13	3.11	3.05	3.42	3.29							
Cv10	3.21	3.36	2.94	2.82	3.01	3.17	3.29	3.23	3.30						
Cont	3.34	3.43	3.13	3.10	3.20	3.27	3.43	3.34	3.24	3.30					
Unw	3.47	3.96	3.22	3.27	3.36	3.45	3.57	3.65	3.87	3.57	3.44				
Cir1	3.34	3.25	3.06	3.27	3.14	3.13	3.02	3.21	3.26	2.92	3.12	3.41			
Cir2	3.58	3.57	3.39	3.34	3.13	3.32	3.57	3.32	3.43	3.50	3.28	3.72	3.38		
Cir3	3.33	3.22	3.03	3.19	3.14	3.19	3.14	3.33	3.26	3.18	3.38	3.57	3.13	3.52	
Apt	3.51	3.34	3.30	3.33	3.19	3.41	3.53	3.44	3.55	3.31	3.22	3.69	3.56	3.37	3.37

5. CONCLUSION

A method of representing speech utterances in feature domain using an AM-FM model has been proposed. Using a dynamic time warping process, features obtained from demodulated amplitude envelope and instantaneous frequency are able to discriminate well between different utterances, independent of the speaker. For a small database of 16 word utterances extracted from a continuous stream of speech, modulation based features around the first four formants showed a recognition rate of 75 percent. This result demonstrates that the proposed modulation based features have a potential to achieve high recognition scores in a keyword spotting system. Further work on a larger database employing time trajectories of the modulation features to increase the difference in dissimilarity scores between utterances of the same word and between different words is in progress.

Acknowledgement: The authors gratefully acknowledge a version of the DTW code made available on the Web by Prof. Dan Ellis, Columbia University, New York.

6. 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] 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.
- [4] 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.
- [5] 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.
- [6] Teager, H.M., and S. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract," *NATO Advanced Study Inst. on Speech Production and Speech Modeling*, Bonas, France, 1989, Kluwer Acad. Pub., 1990.
- [7] P. Maragos, T.F. Quatieri, and J.F. Kaiser, "Speech Nonlinearities, Modulations, and Energy Operators," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (ICASSP '91), pp. 421-424, 1991.
- [8] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (ICASSP '90), pp. 381-384, 1990.
- [9] F. Jabloun, A.E. Cetin and E. Erzin, "Teager Energy based Feature parameters for Speech recognition in Car Noise," *IEEE Signal processing Letters*, vol. 6, No. 10, pp. 259-261, Oct. 1999.
- [10] D. Dimitriadis, P. Maragos and A. Potamianos, "Auditory Teager Energy Cepstrum Coefficients for Robust Speech recognition," *Proc. European Conf. on Speech Communication technology - Interspeech 2005*, Lisbon, Portugal, pp. 3013-3016, Sep. 20, 2005.
- [11] K. Gopalan, "Pitch Estimation Using a Modulation Model of Speech," *Proc. of the International Conference on Signal Processing* (ICSP 2000), World Computer Congress, Beijing, China, Aug. 2000.