NOVEL FEATURE EXTRACTION FOR NOISE ROBUST ASR USING THE AURORA 2 DATABASE

Penny Hix, Stephen Zahorian, Fansheng Meng
phix@odu.edu, szahoria@odu.edu, fmeng@odu.edu

Department of Electrical Engineering
Old Dominion University, Virginia, U.S.A.

ABSTRACT

This paper presents speech signal modeling techniques that are well suited to robust recognition of connected digits in noisy environments. After several preprocessing steps speech is represented by a block-encoding of discrete cosine transform of its spectra. In this paper we combine linear predictive coding (LPC), morphological filtering, and long block lengths to achieve robust features for improved recognition in noisy environments. The spectral envelope is first estimated by LPC. Subsequent morphological filtering enhances the peaks while smoothing the valleys, which are more affected by noise in the signal. These techniques were tested with the Aurora 2 database and the standard HMM recognizer as defined by the ETSI STQ-AURORA DSR Working group for WI007. With no major increase in computational demand a 23% word error rate (WER) reduction has been achieved as compared to the WI007 baseline MFCC front-end for multi-condition training condition. The basic conclusion is that the features resulting from the methods presented here perform better than cepstral features for ASR of noisy speech.

1. INTRODUCTION

Although very high levels of accuracy have been achieved on clean speech, recognition accuracy of ASR systems is seriously degraded when the acoustic conditions are noisy and/or training and testing environments are mismatched. Numerous speech research groups are currently studying various methods to surmount this problem. In order to evaluate and compare the performance of proposed noise robust algorithms, the Aurora 2.0 database was designed [1]. The database was created by first down sampling the TI-digits database from 20 kHz to 8 kHz. In order to simulate realistic telecommunications terminal and equipment characteristics the database was then filtered with the G.712 and MIRS filters, as defined by the International Telecommunications Union (ITU). Eight (8) different real-world noises types were artificially added to the filtered signals.

The ETSI STQ-AURORA DSR Working group standardized the WI007 front-end. 12 Mel frequency cepstral coefficients (MFCCs) are computed using a 25ms frame length with a frame rate of 10ms. The frame level log energy and C_0 term are also computed as additional acoustic features, resulting in a 14-component vector containing static features. Additional details of the WI007 front-end are provided in [1]. Front-end feature extraction evaluation is performed by the standard HMM recognition system, also defined by the Aurora working group.

Feature extraction based on the encoding of global spectral shape has been adopted by the Old Dominion University Speech Communication Lab [2-5]. Our method differs from standard MFCC analysis in that the DCT is applied directly to the log-scaled spectrum. The resulting features are called DCTCs (Discrete Cosine Transform Coefficients). A DCT modified by a bilinear warping function is used to mimic the non-linear speech perception of the human ear [4]. Spectral/temporal features are computed by the Discrete Cosine Series Expansion of the DCTCs. The resulting parameters are called DCSCs (Discrete Cosine Series Coefficients). This technique applies the DCT over a block or stack of frame features. In this work we preprocess the speech signal using linear predictive (LP) coding with morphological filtering (MF) combined with long block lengths. This spectrum is then log-scaled and our standard feature extraction is applied to the log-scaled LP-MF spectrum.

A commonly accepted approach to achieving robustness is acoustic feature compensation and normalization by cepstrum subtraction and cepstrum mean subtraction. Using these methods a word error rate (WER) reduction of 13% has been achieved [6]. On the multi-condition data, over SNR 20 through SNR 0, Macho and Cheng achieved an average WER reduction of 29.90% via SNR-dependent waveform processing [7]. More recently Tsai and Lee achieved an average WER reduction of 9.15%, over all SNR levels, using progressive histogram equalization [8].

2. SPECTRAL ENVELOPE ESTIMATION USING LINEAR PREDICTIVE CODING

Linear Predictive coding analyzes the speech signal using an auto-regressive time-domain model or all-pole spectral model. The linear predictive model has been found to be effective for modeling formants corresponding to resonances of the vocal tract, which carry much speech information. The perceptual relevance of LP, combined with computational efficiency, have made LP analysis very effective for low bit-rate speech coding [10], and also

as an intermediate step for computing speech features for ASR (RASTA-PLP, Hermanksy). Linear prediction can also be used to create an LP inverse filter, which removes the formant information and creates a flat spectrum signal (residual) useful for pitch tracking. The transfer function for linear prediction analysis is given by:

$$H(z) = \frac{G}{\sum_{i=0}^{p} a_i z^{-i}}, \qquad (2.1)$$

where the a_i are linear prediction coefficients, with $a_0 = 1$. When the LP order is suitably chosen the magnitude frequency response estimates the envelope of the signal spectrum. The all pole nature of linear predictive coding produces a spectral envelope with sharp peaks. The accuracy of the spectral envelope will be highly affected by the type of signal analyzed.

3. MORPHOLOGICAL FILTERING

In this work we use morphological filtering to broaden major harmonic peaks and suppress low-amplitude sections of the spectrum. By selecting harmonic peaks and ignoring low level components, which are adversely affected by noise, morphological filtering can be used to reduce noise in the spectrum.

Dilation is a morphological operation that allows objects to expand, thus potentially filling in small holes and connecting disjoint objects. Erosion, a second principle of morphological operation, shrinks objects by eroding their boundaries. A structuring element determines exactly how the objects will be dilated or eroded [9]. In our processing, dilation was used to emphasize and broaden harmonic peaks and to eliminate sharp dips in the spectrum.

4. ALGORITHM DEVELOPMENT

For each spectral frame, *X(f)*, a spectral representation is obtained using the transfer function for linear prediction given in equation 2.1. Due to its all-pole nature sharp peaks are a consequence of LPC processing. In order to select the harmonic peaks morphological filtering is applied to the peaky spectral LPC output spectrum. Figure 4.1 illustrates these steps, depicting the log magnitude spectrum, the spectrum computed from LP coefficients, and the spectrum after morphological filtering. The LP analysis is of order 75 and a morphological filter window width of 250 Hz have been chosen based on values that result in the highest recognition accouracy in experiments with Aurora 2.0, as described in section 5. The peak-based spectral smoothing, which we refer to as LP_MF, is illustrated in Figure 4.1, for a single frame of speech. Our standard feature extraction method is applied to the resulting LP-MF spectrum.

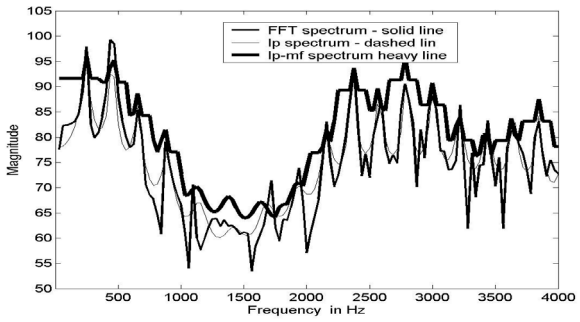

Figure 4.1 spectral processing, showing FFT, LP, LP_MF spectra.

Acoustic features for encoding the speech spectrum are computed as follows:

$$\mathrm{DCTC(i)} = \int_{o}^{1} \log(\mathrm{LP\_MF(f)}) \cos[\pi \mathrm{ig(f)}] \frac{\mathrm{dg}}{\mathrm{df}} \mathrm{df}. \qquad (4.1)$$

DCTC(i) is the *ith* feature as computed from a single spectral frame, *X(f)*. Let

$$\phi_i(\mathrm{f}) = \cos[\pi \mathrm{ig(f)}] \frac{\mathrm{dg}}{\mathrm{df}} \qquad . (4.2)$$

Then we can rewrite equation 4.1 as

$$\mathrm{DCTC(i)} = \int_{o}^{1} \log(\mathrm{LP\_MF(f)})\, \phi_i(\mathrm{f})\, \mathrm{df}. \qquad (4.3)$$

The terms computed with equation 4.3 are equivalent to cepstral coefficients. DCTC parameters are computed using a bilinear warping (g in Eqs. 4.1 and 4.2),

$$\mathrm{f'} = \mathrm{f} + \frac{1}{\pi} \tan^{-1}\left\{ \frac{\alpha \sin(2\pi \mathrm{f})}{1 - \alpha \cos(2\pi \mathrm{f})} \right\}, \alpha = 0.45. \qquad (4.4)$$

In order to encode the trajectory of the short-time spectra DCSC features are computed. The coefficients of the modified cosine expansion over the segment interval represent the trajectory of the DCTCs. This expansion allows non-uniform time resolution as follows. Spectral feature trajectories are encoded by the cosine transform over time using:

$$\mathrm{DCSC(i,\ j)} = \int_{o}^{1} \mathrm{DCT\,C'(i, t')} \cos(\pi\ \mathrm{jt'})\, \mathrm{dt'}. \qquad (4.5)$$

The DCSC(i,j) terms represent both spectral and temporal information over a speech segment. As with DCTCs we can re-express equation 4.5 as follows:

$$\mathrm{DCSC(i, j)} = \int_{o}^{1} \mathrm{DCTC(i, t)} \cos[\pi\ \mathrm{jh(t)}] \frac{\mathrm{dh}}{\mathrm{dt}} \mathrm{dt}. \qquad (4.6)$$

Letting $\theta_j(t) = \cos[\pi\ jh(t)]\ \frac{dh}{dt}$, $h(t)$ a Kaiser window,

$$DCSC(i,\ j) = \int_0^1 DCTC(i,\ t)\ \theta_j(t)\ dt \qquad (4.7)$$

Thus, feature trajectories are represented using the static feature values for each frame, but with varying resolution over a segment consisting of several frames.

5. EXPERIMENTAL RESULTS

All experiments were performed with the HTK toolkit for implementing an HMM configured using the reference parameters provided in [1]. The models are simple left-to-right, no skip, whole word models with 16 states per word (18 states in HTK notation). There are 3 Gaussian mixtures per state with a diagonal covariance matrix. Two (2) pause models are defined. "Sil" models consisting of 3 states with 6 Gaussian mixtures per state model the pause before and after each utterance. "Sp" models pauses between words. It consists of a single state which is tied with the middle state of the "sil" model.

For all experiments reported here analysis was performed using 30 ms frames, with a 10 ms frame spacing, and a frequency range of 0 to 3990 Hz. For each frame, 13 DCTCs were computed, and then encoded with a 3 term DCS expansion over several frames (a block). Overall, the highest accuracy was obtained using a 75th order LP model for each spectral frame, a 250 Hz window width for morphological filtering, and 11 frames per block (i.e., approximately 110 ms block). However, in this section, three sets of experiments are reported, to illustrate the effects of varying the number of frames per block, the LP order with no morphological filtering, and varying the LP order with morphological filtering.

Experiment 1: Effects of LP order

In this experiment we study the effects of linear prediction with no morphological filter. The number of frames per block was fixed to 11, and the linear predictive filter order was varied. Training was performed on Aurora 2.0 multi-condition data, and testing with test sets A, B, and C of the same database. Table 1 gives results for linear predictive orders of 0, 15, 25, 50, 75, and 100 coefficients. Note that for the case of LP order zero, linear prediction was not used.

Results of experiment 1:

Table 1: Word Accuracy for varying LP orders with no morphological filtering

Test Set / LP order	A	B	C	Average
0	83.98	79.52	72.87	79.97
15	89.57	85.52	85.40	**87.12**
25	89.04	84.51	85.02	86.42
50	85.55	81.58	77.67	82.38
75	84.97	80.07	73.35	80.69
100	85.02	80.07	73.45	80.73

Conclusion of experiment 1: The results show that with 15 terms LP based DCTC and DCSC features are slightly more noise robust than those produced by the baseline (WI007).

Experiment 2: In this experiment we used 11 frames per block, and varied the order of the linear prediction. This experiment differs from experiment one in that we now included morphological filtering with a window width of 250 Hz. Training was again performed on Aurora 2.0 multi-condition data, and testing with test sets A, B, and C of the same database.

Results of experiment 2:

Table 2: Word Accuracy for varying LP orders with morphological filtering window width of 250 Hz

Test set / LP order	A	B	C	Average
0	90.68	87.14	88.71	88.87
15	89.17	85.46	85.14	86.88
25	89.93	85.92	87.44	87.83
50	90.59	86.57	88.30	88.52
75	90.82	87.73	89.04	**89.23**
100	90.56	86.69	89.16	88.73

Conclusion of experiment 2: A 23% WER reduction was achieved with LP_MF features as compared to those obtained with baseline WI007. All results with morphological filtering are superior to those without morphological filtering, except for an LP order of 15, an order closer to that more typically used.

Experiment 3: In this experiment a 75th order LPC model and a morphological window width of 250 Hz were used. The block length was varied from 3 frames to 19 frames in steps of 2 (or approximately 30 ms to 190 ms, in steps of 20 ms). Training was performed on Aurora 2.0 multi-condition data, and testing with test sets A, B, and C of the same database.

Results of experiment 3:

Table 3: Word Accuracy as a function of block length

Test Set / Block Length	A	B	C	Average
3	86.98	81.92	81.30	86.34
5	88.95	84.53	84.74	83.82
7	90.38	85.83	87.43	87.97
9	90.45	86.23	88.44	88.45
11	90.82	87.73	89.04	89.23
13	90.04	86.72	89.08	88.52
15	88.30	85.02	86.60	86.65
17	87.84	84.61	86.00	86.18
19	87.05	83.59	84.75	85.21

Conclusion of experiment 3: LP_MF features computed with block length of 11 resulted in the highest overall accuracy. The word error rate is 18.4% lower than results obtained with a block length of 3 frames, and 33.4% lower than results obtained with a block length of 5. Block lengths longer than 13 frames (130 ms) result in degraded performance.

Conclusions and future work: We have presented a novel combination of front end signal processing consisting of linear prediction with morphological filtering for spectral envelope smoothing, followed by discrete cosine analysis in frequency to compute spectral features, followed by a second discrete cosine transform of spectral features over time to encode trajectory information. The combinations of LP_MF DCTC features, and the dynamic features computed from them with a long time window, produce a 23% WER reduction on the Aurora 2.0 multi-condition training data. The biggest differences between the processing reported in this paper, and more typical processing, are the use of a much higher LP order (75^{th} order versus more typical order of 15), use of morphological filtering, and use of much longer block lengths (11 frames per block versus more typical use of 3 or 5 frames per block). The original motivation for using morphological dilation was to remove the low energy (and presumably noisy) spectral intervals between F_0 harmonics in the original spectrum. It is unclear as to why the very high order LP model combined with morphological filtering (which smoothes the spectrum) results in noise robust features.

Table 4 gives a comparison of our results with the original WI007 front-end, the updated version of WI007, and results produced by Evans & Mason using QBNL and NLSS based features. Using the original, less robust, recognizer our features performed quite reasonably when compared to results produced by the updated recognizer. We are currently testing our features on the Aurora 3.0 front-end and will be using the updated version of the HMM recognizer defined by the WI007 update. In addition, we will begin testing a new feature set defined by morphological filtering followed by linear prediction (MF-LP features) because the peak selection property of morphological filtering along with the peak sharpening property of linear prediction should further enhance robustness in noisy conditions.

Table 4: Comparison of results

Test Set	A	B	C	Average
WI007 baseline	87.81	86.27	83.77	85.95
Evans & Mason QBNL&NLSS	90.00	88.58	88.05	88.60
LP_MF	90.82	87.73	89.04	89.23
WI007 update (20 Gaussian mixtures)	92.5	90.47	90.75	90.26

6. REFERENCES

[1] David Pearce, Hans-Gunter Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions", ICSLP, October 16, 2000

[2] M. Karnjanadecha, S.A. Zahorian, "Robust Feature Extraction for Alphabet Recognition," *Proc. ICSLP 98*, Sydney, Australia, vol. 2, pp. 337-340, 1998.

[3] M. Karnjanadecha and S.A. Zahorian, "Signal Modeling for Isolated Word Recognition," *Proc. ICASSP 99*, vol. 1, pp. 293-296, Phoenix, AZ., March 1999.

[4] Zahorian, S. and Jagharghi, A. "Spectral-shape Features versus Formants as Acoustic Correlates for Vowels," *J. Acoust. Soc. Amer.*, vol. 94, pp 1966-1982, 1992.

[5] Zahorian, S. A., Silsbee, P. L., and Wang, X., "Phone Classification with Segmental Features and a Binary-Pair Partitioned Neural Network Classifier," *Proc. ICASSP 97*, pp. 1011-1014, Munich, Germany, April 1997.

[6] H.K. Kim, R.C. Rose, "Evaluation of Robust Speech Recognition Algorithms for Distributed Speech Recognition in a Noisy Automobile Environment," Proc. Int'l Conf. on Spoken Language Processing (ICSLP), pp. 233-236, 2002

[7] Dusan Macho and Yan Ming Cheng, "SNR-Dependent Waveform Processing for Improving the Robustness of ASR Front-end," Proc. Int'l Conf on Acoustics, Speech, and Signal Processing, pp 305-308, 2001

[8] Sang-nien Tsai and Lin-shan Lee, "A New Feature Extraction Front-end for Robust Speech Recognition using Progressive Histogram Equalization and Multi-eigenvector Temporal Filtering," Proc. Int'l Conf on Spoken Language Processing, October 2004

[9] Hyun-Soo Kim and W. Harvey Holmes, "Spectral Estimation and Speech Analysis Techniques Using Morphological Filters"

[10] Editorial Board: Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, Victor Zue "Survey of the State of the Art in Human Language Technology," National Science Foundation & European Commission, 1996

7. ACKNOWLEDGEMENT

This work was partially supported by NASA NGT-1 01029, for the first author of this paper under the GSRP program.