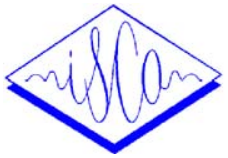

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

COMPENSATION FOR ENVIRONMENTAL AND SPEAKER VARIABILITY BY NORMALIZATION OF POLE LOCATIONS

Juan M. Huerta and Richard M. Stern

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

ABSTRACT

We present a compensation technique that corrects for the effects of noise and variability of speaker and environment on speech recognition accuracy by modifying the positions of the poles representing the speech signal in the z-plane. This modification yields pole locations with statistics that more closely match the statistics of the distribution of clean training speech. The parameters of the mapping are obtained from statistics of the distribution of the poles of the training and testing speech. Compensation is performed by direct modification of both the angle and the radius of pole locations, and also by evaluating the cepstrum along a circle of radius less than 1 in the z-plane to enhance the salience of spectral peaks. These procedures are evaluated using the DARPA Resource Management database using added white noise. They are shown to compensate for the effects of environmental degradation, particularly at low SNRs.

1. INTRODUCTION

The performance of automatic speech recognition systems depends greatly on the extent to which their training and testing data are acoustically matched. A mismatch between training and testing data can be a consequence of differences in the acoustical environment and differences between speakers' articulatory characteristics. Acoustical compensation algorithms aim to reduce these mismatches.

In this paper we consider the effects of different sources of acoustic variation on the locations of poles of speech in the z-plane. We present some observations about the behavior of these pole locations as the acoustical environment is modified. Based on these observations, we propose several methods that normalize the location of the poles of noisy speech so that difference between their long-term statistical descriptions and those of poles representing clean reference speech are minimized.

Several previous researchers have described the effects of additive noise on the AR parameters of a speech signal and have attempted to compensate for the effects of additive noise (*e.g.* [4, 5, 6]). For the specific case of additive white Gaussian noise on speech, this effect has been modeled in [6] as a displacement of the poles representing speech towards the origin of the z-plane. Such pole displacement is in accordance with the notion that additive noise causes the norm of the cepstral vector to shrink and the spectrum to flatten [6].

Another source of mismatch between training and testing data is that of variability in vocal tract anatomy across speakers. This variation is reflected mostly in the locations of the spectral resonances of speech (*i.e.* the formant locations). Some speaker normalization techniques that have been proposed (*e.g.* [8, 10]) warp the frequency axis so that the locations of spectral peaks match more closely the locations of spectral peaks from speakers in the training data. Vocal tract variation is reflected primarily by a displacement of the means of the distributions of the angles of the poles. In [9] an algorithm was proposed in which the mapping is performed directly on the poles of the speech.

When considered together, the effects of additive noise and speaker variation result in a distortion of both the angles and radii of the locations in the z-plane of the poles representing speech. While the distortions of radius and angle have previously largely been considered separately (*e.g.* [4, 8]) they are actually components of a general mismatch based on differences in environment and speaker. These effects are hard to describe analytically due to the complex nonlinear relation that exists between the pole locations and the speech signal. Estimating the parameters of the environment based on the locations of the poles and restoring the original signal based on an analytical model of these effects is at best, a challenging task.

In this paper, we reduce the mismatch introduced by the environment by directly modifying the locations of the poles that represent the noisy signal. Our approach differs from the previous methods in that we consider both the radius and angle of the pole locations and we utilize the statistical distributions of the poles of the degraded and clean speech to reduce the differences between these distributions. Since no analytical model of environment or speaker is employed, these methods are expected to be effective under different environmental phenomena.

2. EFFECTS OF NOISE ON POLE LOCATIONS

To illustrate the effects of additive noise on pole locations and distributions, we contaminated a short segment of voiced speech with additive white noise at different signal-to-noise ratios. Figure 1 shows how the poles of the original clean speech are displaced to different loci as signal-to-noise ratio (SNR) is decreased. For this particular realization of the noise and speech, the locations of each of the complex poles seem to follow a continuous trajectory from their initial to their final position for the SNRs considered. (This is not necessarily the case for any two additive signals.) The effects of the noise on the pole locations are nonlinear.

In Figure 2 we illustrate how noise affects the statistics of pole locations in the z-plane. The first row shows the distribution of the radius and angle for the poles corresponding to the second formant under clean conditions for 15 seconds of speech. The second and third rows show the histograms of the angle and radius of the first pole of the same segment of speech with noise added at 5 dB and 10 dB. As is evident from the graphs, the distributions of the poles on the z-plane are influenced by the intensity and spectral shape of the noise.

The distributions of the angle and the radius show clearly different shapes. The distributions of the angles are more symmetric, and might be reasonably approximated by a Gaussian curve. The effects of noise on this distribution can be represented as changes in the mean and variance of the distributions.

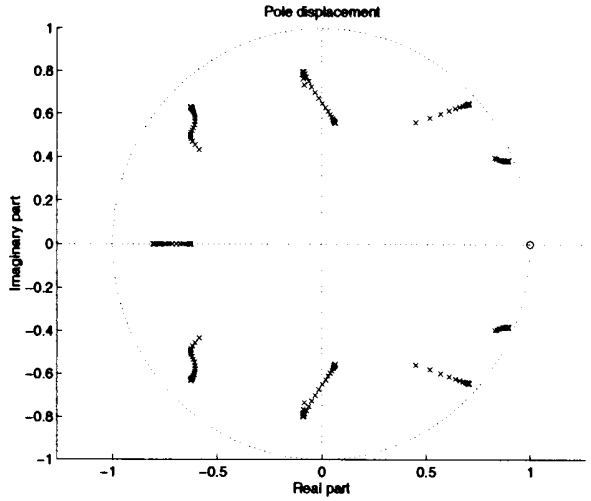

Figure 1 Pole locations of speech for various values of SNR.

The distributions of the radii, however, are non-symmetric, with modes that are close to the unit circle, and they cannot be reasonably characterized as Gaussian. Additive noise affects these distributions by displacing the modes towards the origin as well as by increasing the width of the mode (*i.e.*, it "smears" this distribution).

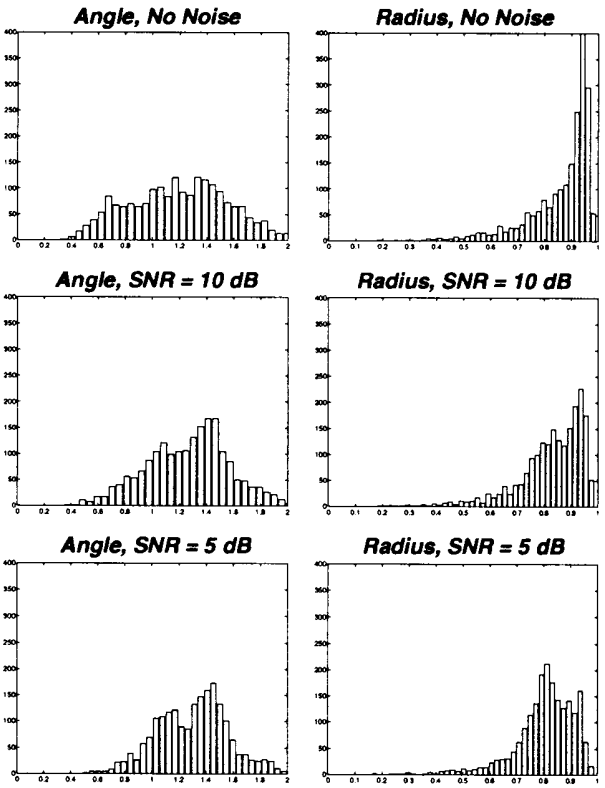

Figure 2 Histograms of the angle and radius of the poles corresponding to the second formant at three SNRs.

The angles and radii of the poles of speech derived from linear prediction associated with a given formant, occur in a non-independent fashion. Ideally, one would like to compensate theses poles considering the joint distribution of angle and radius. This is hard to do in a parametric way, because of the poor fit of a multivariate Gaussian model to the statistics of the poles. For simplicity, we treat the effects of angular and radial displacement independently in the work described in this paper.

3. SPEECH COMPENSATION BY POLE NORMALIZATION

We describe three techniques that map the poles representing speech in the testing data to a new set of locations that better match the corresponding locations from the clean training data. The statistics of normalized clean speakers are computed by averaging over a long sample of clean speech recorded from various training speakers. The distributions of the incoming speech to be normalized are generally estimated over a smaller amount of speaker-specific speech.

3.1. Angular Compensation

Angular compensation corrects the angular positions of the poles using statistics of the clean and noisy speech. This statistical information is collected in terms of formants, which are assumed to be represented by the four sets of complex poles with the angles of smallest magnitude, regardless of the corresponding radii. The mean and variance of each formant distribution are calculated.

Remapping of the angles of pole locations is performed by determining the relative angular position of each complex pole pair with respect to the immediate upper and immediate lower mean values of the formants of the test speaker. Using linear interpolation, this relative angular position is mapped to a new value between the immediate upper and immediate lower mean values of the formants of a normalized clean speakers. For example, if the angle of a specific pole falls between the first and second formant of the active speaker, the normalized angle value would be the angle that is in a corresponding location between the first and second formant of the normalized speaker. Specifically, the angle θ_i is mapped to the angle θ'_i according to the following expression

$$\theta'_i = \bar{g}_k + (\theta_i - \bar{f}_k)\frac{(\bar{g}_{k+1} - \bar{g}_k)}{(\bar{f}_{k+1} - \bar{f}_k)}$$

where $\bar{g}_k$ and $\bar{g}_{k+1}$ represent the mean values of the formants k and $k+1$ of the clean speech $\bar{f}_k$ and $\bar{f}_{k+1}$ represent the mean values of the formants k and $k+1$ of the noisy speech. The angle θ'_i falls in the frequency interval delimited by $\bar{f}_k$ and $\bar{f}_{k+1}$. This mapping can be interpreted as a piecewise linear warping function of the locations of the poles where the breakpoints of the function are the means of the formants for the clean and noisy data.

3.2. Radial Compensation

Remapping of the radii is accomplished by associating each pole with a formant and then applying the following linear mapping:

$$r'_i = a_i r_i + b_i$$

Each a_i and b_i are calculated for every formant by performing a linear regression between the endpoints of 20 equally-spaced percentile values of the histograms of the radii of the poles of the noisy and clean speech. In this way, the histograms of the resulting mapped poles will be similar to the poles of the clean speech. The poles resulting from this mapping, however, might not fall inside the unit circle. Precautions need to be taken such that this doesn't occur.

3.3. Enhancing the Spectrum: Off-axis evaluation of the Cepstral Vector

A very simple approach that enhances the spectral peaks of the

speech is accomplished by evaluating the z-transform closer to the locations of the poles. As noted in [6], by adding noise to the speech signal, the poles of the signal are displaced towards the origin. The term *off-axis spectrum* [7], refers to an enhanced spectrum obtained by evaluating the z-transform along a circle in the z-plane of radius less than one. Calculating the cepstral vector using poles enhanced with this method is straight forward. Letting $A(z)$ be the LPC polynomial of the signal and $r < 1$ be the new radius along which the spectrum will be evaluated,

$$A(rz) = \sum_{i=0}^{M} a_i (rz)^{-i} = \sum_{i=0}^{M} (a_i r^{-i}) z^{-i}$$

If we derive the cepstral coefficients by replacing a_i by $a_i r^{-i}$ in the cepstral recursion, these cepstral coefficients will correspond to those of a spectrally enhanced signal. The value of r is determined by ratio of the average of the radii of all the poles representing the degraded speech in the testing domain divided by the corresponding average of the radii of the poles from the clean speech used to train the system. Particular care should be taken when implementing this enhancement routine: if r is smaller than the magnitude of the largest pole, then the system represented by the polynomial will be unstable. This method can be extended by evaluating the cepstral vector along a non-uniform, non-circular contours as in the chirp z-transform [3].

4. POLE NORMALIZED SPEECH FEATURES

We refer to the acoustic features used in these experiments as Mel Scale based Linear Prediction Cepstral Coefficients (MLPCC). The procedure to obtain the MLPCC vectors is illustrated in Figure 3. Speech is sampled at 16 kHz, windowed using a 20-ms Hamming window, and the 512-point DFT of the windowed speech is calculated for each frame. 11 Mel-scale based autocorrelation coefficients are obtained by computing the inverse DCT of the squared magnitude of the DFT coefficients after the conventional Mel-scale triangular weighting functions. (The Mel-scale based autocorrelation provides robustness in noisy conditions and was inspired by [2]). An 11^{th}-order LPC polynomial is obtained from the autocorrelation coefficients using Levinson-Durbin recursion, from which the pole locations are obtained.

The poles are then compensated using one of the algorithms described in Section 3. After the set of normalized poles is obtained it is necessary to verify that the new poles actually exhibit the set of desired properties described in Section 4.1 below. The LPC vector is recalculated using the normalized pole locations, and the spectral enhancement based on evaluation inside the unit circle is applied as described in Section 3.3. Finally, a 13^{th}-order cepstral vector is recursively obtained from the new LPC vector, and the corresponding delta and delta-delta coefficients are calculated.

4.1. Practical Considerations

Altering the positions of the poles according to the procedures described in Section 3 can result in normalized pole locations that are invalid. Specifically, special care must be taken regarding the following phenomena:

- **Angle crossing π:** Angular correction might produce normalized poles whose angle exceed the value of π in the z-plane. This corresponds to a form of aliasing and must be avoided.
- **Radius value exceeding 1:** Radial correction might produce poles falling outside the unit circle, which normally would result in a unstable system.
- **Formant swaps:** Angular compensation algorithms might change the angular ordering of the poles, which is equivalent to inverting the order of the formants.
- **Instabilities from evaluating the spectrum inside the unit circle:** Shrinking the locus of evaluation of the LPC polynomial can also leave poles outside the new circle.

A strategy for detecting and correcting the above possibilities must be devised. The simple alternative used in the present experiments was to discard each frame containing a problematic pole location, duplicating the pole locations of the previous frame. Some mapping algorithms, such as the angular compensation routine, will always map the poles to valid locations.

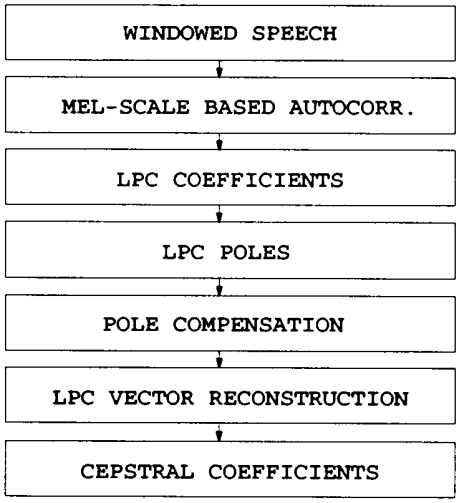

Figure 3 Block diagram of the MLPCC front end with pole normalization.

5. EXPERIMENTAL RESULTS

Experiments were conducted using the Speaker Independent set of the speaker-independent portion of the ARPA Resource Management RM1 Task. Figure 4 shows the results of these experiments. In every situation, acoustic models were trained using clean speech that was compensated using the statistics of the complete training set. Tests were performed using four different SNRs: 5 dB, 10 dB, 20 dB, and no added noise (clean speech). Results are shown when only the angle, only the radius, and both the radius and angle are compensated. Uncompensated MLPCC and MFCC results are shown for comparison.

Results using off-axis cepstral evaluation were also obtained under similar noise conditions. In this case acoustic models were trained from clean speech without any sort of compensation. The optimal value of r was determined for every utterance as described in Section 3.3. The results obtained with this method are shown in Table 1 along with the MLPCC baseline. Comparing the results obtained using radius-only normalization (Fig. 4) with those using off-axis cepstral evaluation (Table 1) the superiority of the radius normalization approach over off-axis evaluation of the cepstra is evident. This difference in performance is probably due to the fact that with radius-only compensation every pole location is displaced depending on its relative position with respect to the other poles and to the percentiles of the distributions of the radii. In the off-axis evalua-

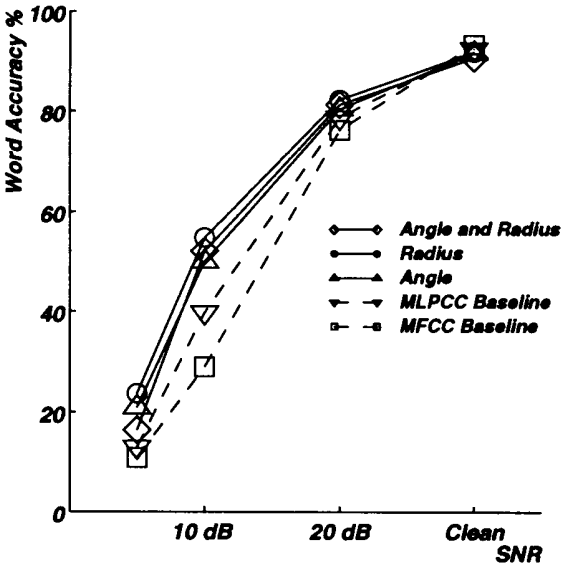

Figure 4 Recognition accuracy on the RM1 task using pole compensation.

tion technique every pole is mapped using the same mapping parameter, r. While radius compensation offers greater robustness to noise, it requires the explicit calculation of the poles in every frame. Off-axis evaluation of cepstra, on the other hand, is trivial to implement as it operates exclusively on the LPC vector coefficients.

	5 dB	10 dB	20 dB
MLPCC	13.0%	39.8%	78.7%
Off-axis	14.9%	42.5%	78.7%

Table 1. Recognition accuracy using off-axis cepstral evaluation for the RM1 task.

6. DISCUSSION AND SUMMARY

We presented a set of algorithms that reduce the statistical mismatch introduced by the environment between testing and training data. These techniques operate directly on the poles of the speech and do so based solely on their statistical descriptions.

It was found that compensation of the radii and angles of the poles, either individually or i4n consort, provided substantial improvements in recognition accuracy at lower SNRs. Neither approach improved the recognition accuracy of clean speech, despite the successful application of frequency-warping approaches to speaker normalization in previous work (*e.g.* [1, 8, 10]). Evaluation of the frequency response along a circle in the z-plane of radius less than 1 provided modest improvement at lower SNRs, but not nearly as much as full normalization of the radii and angles of the pole locations.

ACKNOWLEDGMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.. Juan M. Huerta is thankful to CONACYT, México, for its generous support.

7. REFERENCES

[1] E. Eide and H. Gish "A Parametric Approach to Vocal Tract Length Normalization", Proc. ICASSP 1996, Vol. 1, pp. 339-341 .

[2] H. Hermansky "Perceptual Linear Predictive (PLP) Analysis of Speech", J. Acoust. Soc. Amer., Vol. 87, pp. 1738-1752, 1990.

[3] J. M. Kates "An Auditory Spectral Analysis Model Using the Chirp z-Transform", IEEE Trans. on ASSP, Vol. 31, pp. 148-156, 1983.

[4] S. M. Kay "Noise Compensation for Autoregressive Spectral Estimates", IEEE Trans. on ASSP Vol. 28, pp. 292-303, 1980.

[5] L. Lee and R. Rose, "Speaker Normalization Using Efficient Frequency Warping Procedures", Proc. ICASSP 1996, Vol. 1, pp. 353-356.

[6] J. S. Lim and A. V. Oppenheim "All-Pole Modeling of Degraded Speech", IEEE Trans. on ASSP Vol. 26, pp. 197-210, 1978.

[7] D. Mansour and B.-H. Juang "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition", IEEE Transactions on ASSP Vol. 37, pp. 1659-1671, 1989.

[8] J. D.Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag New York, 1976.

[9] J. McDonough, G. Zavaliagkos and H. Gish, "An approach to Speaker Adaptation Based on Analytic Functions", Proc. ICASSP 1996, Vol. 2, pp. 721-724.

[10] S. Wegmann, D. McAllaster, J. Orloff and B. Peskin "Speaker Normalization on Conversational Telephone Speech", Proc. ICASSP 1996, Vol. 1, pp. 339-341.