ROBUST SPEECH RECOGNITION BY NORMALIZATION OF THE ACOUSTIC SPACE

Alejandro Acero and Richard M. Stern

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

Abstract

In this paper we present several algorithms that increase the robustness of SPHINX, the CMU continuous-speech speaker-independent recognition system, by normalizing the acoustic space via minimization of the overall VQ distortion. We propose an affine transformation of the cepstrum in which a matrix multiplication performs frequency normalization and a vector addition attempts environment normalization. The algorithms for environment normalization are very efficient and they improve dramatically the recognition accuracy when the system is tested on a microphone other from the one on which it was trained. The frequency normalization algorithm applies a different warping of the frequency axis to different speakers and it achieves a 10% decrease in error rate.

1. Introduction

Building spoken language systems is difficult due to the enormous variability present in the speech signal. Although a model for the specific variability is often desired, in many cases we have to be content with some kind of "multi-style" training, including data representing most possible conditions. In this paper we describe some algorithms that attempt to increase the system robustness by applying an affine transformation on the cepstrum that normalize the acoustic space.

We are concerned in this study in reducing the long-term variabilities caused by different speakers and acoustical environments. Most current recognition systems are very fragile when taken outside the laboratory into the real world, because they operate under different conditions from those for which they were trained. Especially harmful are the presence of additive noise and spectral tilt. Boll [1] proposed spectral subtraction techniques that with some modifications are still valid today. Some authors (*e.g.* Van Compernolle [2]) propose the use of a microphone array to create a directionality pattern that effectively increases the SNR by reducing noise from undesired directions. While previous approaches are effective in suppressing additive noise, they do not combat distortion introduced by linear filtering (spectral tilt). Stockham [3] proposed blind deconvolution to compensate for these linear distortions. Erell and Weintraub [4] demonstrated improved performance by compensating independently for the effects of noise and spectral tilt. Techniques based on auditory models (*e.g.* Seneff [5]) are also very promising, but they incur a substantial computational burden.

In [6], we presented two algorithms: SDCN and CDCN. SDCN applies a fixed additive correction that depends exclusively on the instantaneous SNR of the input. CDCN first estimates the environmental parameters representing additive noise and spectral tilt using EM techniques, and then performs the appropriate correction. While SDCN is simple and effective, CDCN is more complex but performs better.

In this paper we present two new algorithms to normalize the acoustic space in the cepstral domain, the parameter space of SPHINX. These algorithms, called *Interpolated SNR-Dependent Cepstral Normalization* (ISDCN) and *Fixed CDCN* (FCDCN), are extensions of the SDCN and CDCN algorithms presented in [6]. FCDCN is more computationally efficient than CDCN and at least as effective, although it requires environment-dependent training. ISDCN is also computationally efficient and doesn't require environment-dependent data, but it is not as accurate as CDCN. We also propose a novel method for frequency normalization that increases the recognition accuracy by about 10%, by removing some of the variability of different speakers.

2. Environment Normalization

To accomplish the normalization of the acoustic space we propose the following affine transformation

$$\hat{\mathbf{x}}_i = \mathbf{L}\mathbf{z}_i + \mathbf{w} \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the normalized and unnormalized cepstrum vectors, $\mathbf{w}$ is the environmental correction and $\mathbf{L}$ is the frequency normalization matrix. In ISDCN the correction vectors $\mathbf{w}$ are a function of the instantaneous signal-to-noise ratio (SNR) of the noisy input and the environmental parameters noise $\mathbf{n}$ and equalization $\mathbf{q}$. In FCDCN the correction vectors $\mathbf{w}$ depend on the identity of the closest VQ codeword, as well as the instantaneous SNR.

2.1. Interpolated SDCN

One of the deficiencies of the SDCN algorithm presented in [6] is its inability to adapt to new environments because its correction vectors are pre-computed by comparing cepstra representing simultaneously-recorded speech from the training and testing environments. The ISDCN algorithm can re-estimate these correction from the testing data as it arrives.

In both SDCN and ISDCN, the compensated vector $\hat{\mathbf{x}}_i$ is of the form

$$\hat{\mathbf{x}}_i = \mathbf{z}_i - \mathbf{w}(\mathbf{n}, \mathbf{q}, SNR_i) \quad (2)$$

where **n** and **q** are environmental parameters representing the effects of additive noise and distortion from linear filtering, respectively. In ISDCN the correction vector is expressed as

$$\mathbf{w}_i(\mathbf{n}, \mathbf{q}, SNR) = \mathbf{n} + (\mathbf{q} - \mathbf{n}) f(SNR_i) \quad (3)$$

where the function *f interpolates* between the noise **n** at low SNR and the equalization vector **q** at high SNR, so that the correction vector performs noise suppression at low SNR and equalization at high SNR. We selected *f* to be the sigmoid function

$$f_i(x) = 1 / [1 + \exp(-\alpha_i x + \beta_i)] \quad \alpha > 0 \quad (4)$$

because it satisfies the asymptotic behavior of being $f \approx 0$ at low SNR and $f \approx 1$ at high SNR. It is also monotonic and very smooth.

The noise vector **n** can be reliably estimated by averaging a number of noise frames, as described in [6]. We show in [7] that the equalization vector can be obtained by an EM algorithm that minimizes the accumulated VQ distortion:

1. Start with an initial estimate for $\hat{\mathbf{q}}^{(0)}$ and $j = 1$
2. Label all frames, *i.e.* find the value of $k_i^{(j)}$ that minimizes the VQ distortion
3. Estimate $\mathbf{q}^{(j)}$ from all the frames in the utterance:

$$\hat{\mathbf{q}}^{(j)} = \mathbf{n} + \frac{\sum_{i=0}^{N-1} (\mathbf{z}_i - \mathbf{n} - \mathbf{c}[k_i^{(j)}]) f(SNR_i)}{\sum_{i=0}^{N-1} f^2(SNR_i)} \quad (5)$$

4. If convergence has been reached, stop; else go to step 2.

Since SPHINX's cepstrum codebook does not contain the zeroth order term **z**[0], the value of **q**[0] cannot be computed by Eq. (5). The constraint we used to estimate **q**[0], the gain control, was that the dynamic range of the utterance ($\max\{\hat{\mathbf{x}}_i[0]\} - \mathbf{n}[0]$) had to be constant.

For evaluation, α_i and β_i were set empirically to 3.0 for $i > 0$ and to 6.0 for $i = 0$. The equalization estimate $\hat{\mathbf{q}}$ given by Eq. (5) exhibited a large variance for short utterances which introduced noise into the system. To ameliorate this problem, we only reestimate the first 4 cepstral coefficients of **q**, setting the high order ones to zero. This reflects the fact that the equalization function must be spectrally smooth. Since SPHINX already performs vector quantization, ISDCN can be implemented with little computational overhead.

In Table 1 we compare the performance of the census database [6] with the ISDCN algorithm. These comparisons were obtained using testing data from two microphones: the close-talking Sennheiser HMD224 (CLSTK), and the omnidirectional desktop Crown PZM6FS (CRPZM). In each case the system was trained on processed speech from the CLSTK microphone.

A second evaluation with speech from four microphones is presented in Table 2. In addition to new speech using the CRPZM, this evaluation includes speech from a cardioid desktop Crown PCC160 (CRPCC), a handheld dynamic cardioid Sennheiser 518 (SE518), and an electret supercardioid Sennheiser ME80 (SEME80). The system was trained with processed speech from the CLSTK in all cases. The entries in the table compare the performance of the system with no processing (to the left of the slash) and with environmental normalization using the ISDCN and other algorithms (after the slash). We see that the ISDCN algorithm achieves improved recognition accuracy, although not as much as the CDCN for recordings with low SNR.

TEST	CLSTK	CRPZM
BASE	85.3%	18.6%
SDCN	N/A	67.2%
CDCN	85.3%	74.9%
ISDCN	84.8%	62.1%
FCDCN	N/A	73.1%

Table 1: Performance of the ISDCN and FCDCN algorithms compared with the baseline, SDCN, and CDCN, using testing data from two microphones. The system was trained using processed speech from the CLSTK microphone.

TEST	CRPZM	CRPCC	SE518	SEME80
BASE	84.8/41.8	82.4/70.2	87.2/84.5	83.7/71.4
CDCN	83.3/73.9	81.0/78.5	82.2/83.3	81.5/80.7
ISDCN	86.1/73.7	82.3/75.4	87.2/83.5	83.2/78.5
FCDCN	NA/79.3	NA/77.1	NA/83.4	NA/81.1

Table 2: Analysis of performance of SPHINX for the baseline and the CDCN, ISDCN and FCDCN algorithms using four different microphones (see text).

2.2. Fixed CDCN

In this section we describe the Fixed CDCN algorithm, which combines some of the attractive features from both the SDCN and CDCN algorithms. The motivation for this algorithm is to obtain an algorithm that is as accurate as CDCN and as computationally efficient as SDCN.

The Fixed CDCN applies a correction that depends on the instantaneous SNR of the input (like SDCN), but that is also different for every codeword (like CDCN):

$$\hat{\mathbf{x}} = \mathbf{z} + \mathbf{r}[k, SNR] \quad (6)$$

The selection of the appropriate codeword is done at the VQ stage, so that label *k* is chosen to minimize

$$\|\mathbf{z} + \mathbf{r}[k, SNR] - \mathbf{c}[k]\|^2 \quad (7)$$

This technique has been applied to speech from the desktop CRPZM when the system was trained using the CLSTK close-talking microphone. The new correction vectors were estimated with an EM algorithm that maximizes the likelihood of the data.

The probability density function of **x** is assumed to be a mixture of Gaussian densities as in [6].

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} P[k]\ N_{\mathbf{x}}(\mathbf{c}[k], \Sigma_k) \tag{8}$$

The cepstra of the corrupted speech are modeled as Gaussian random vectors, whose variance depends also on the instantaneous SNR, l, of the input.

$$p(\mathbf{z}\,|\,k, \mathbf{r}, l) = \frac{C'}{\sigma[l]}\ \exp\left(-\frac{1}{2\sigma^2[l]}\,\|\mathbf{z} + \mathbf{r}[k, l] - \mathbf{c}[k]\|^2\right) \tag{9}$$

In [7] we show that the solution to the EM algorithm is the following iterative algorithm. In practice, convergence is reached after 2 or 3 iterations if we choose the initial values of the correction vectors to be the ones specified by the SDCN algorithm.

1. Assume initial values for $\mathbf{r}'[k, l_k]$ and $\sigma^2[l]$.
2. **Estimate** $f_i[k]$, the *a posteriori* probabilities of the mixture components given the correction vectors $\mathbf{r}'[k, l_k]$, variances $\sigma^2[l]$ and codebook vectors $\mathbf{c}[k]$

$$f_i[k] = \frac{\exp\left(-\frac{1}{2\sigma^2[l_k]}\,\|\mathbf{z}_i + \mathbf{r}'[k, l_k] - \mathbf{c}[k]\|^2\right)}{\sum_{p=0}^{K-1} \exp\left(-\frac{1}{2\sigma^2[l_p]}\,\|\mathbf{z}_i + \mathbf{r}'[p, l_p] - \mathbf{c}[p]\|^2\right)} \tag{10}$$

3. **Maximize** the likelihood of the complete data by obtaining new estimates for the correction vectors $\mathbf{r}[k, l_k]$ and corresponding $\sigma[l]$:

$$\mathbf{r}[k, l] = \frac{\sum_{i=0}^{N-1} (\mathbf{x}_i - \mathbf{z}_i)\, f_i[k]\ \delta[SNR_i - l\Delta_{SNR}]}{\sum_{i=0}^{N-1} f_i[k]\ \delta[SNR_i - l\Delta_{SNR}]} \tag{11}$$

$$\sigma^2[l] = \frac{\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} \|\mathbf{x}_i - \mathbf{z}_i - \mathbf{r}[k, l]\|^2\, f_i[k]\ \delta[\,SNR_i - l\Delta_{SNR}]}{\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} f_i[k]\ \delta[\,SNR_i - l\Delta_{SNR}]} \tag{12}$$

4. Stop if convergence has been reached, otherwise go to step 2.

Figure 1 shows the resulting variances $\sigma^2[l]$ obtained after the process for $\Delta_{SNR} = 1\,dB$. The large variance exhibited at low SNR reflects the higher uncertainty in the value of the CLSTK speech given the CRPZM speech that occurs at low SNRs.

We also tried estimating the correction vectors by replacing the sum in Eq. (8) by a maximum. The resulting Eqs. (11) and (12) are still valid, but the *a posteriori* probabilities $f_i[k]$ are now a Dirac function $\delta[k]$, being 1 if k is the VQ label for frame i and 0 otherwise. The recognition rate for the CRPZM, 72.6%, is essentially the same obtained with the previous estimation method. One of the differences between our algorithm and the one suggested by Gish *et al.* [8], is that they made the approximation of maximum by sum in the transformed noisy space, instead of in the clean space. In their algorithm, many different vectors in the clean space are transformed into essentially the same vector in the noisy space at low SNRs. With their method, small fluctuations around the observed vector in the noisy space yield very different labels in the clean space at low SNR, and the restored vectors exhibit the *musical* noise characteristic of spectral subtraction techniques (*e.g.* [7]).

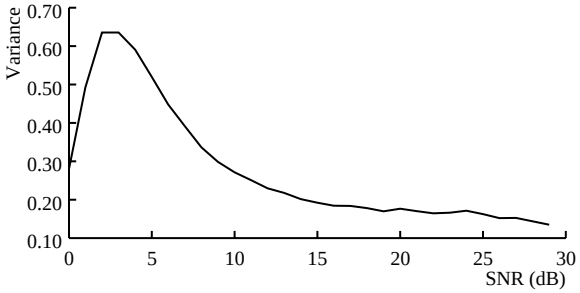

Figure 1: Variance of the difference vector between the CLSTK and the restored CRPZM speech for different input SNR of the CRPZM.

The computational complexity of this fixed CDCN is very low because the correction vectors are precomputed, and it is at least as accurate as CDCN. However, it does require simultaneously-recorded data from the training and testing environments.

3. Frequency Normalization

Speaker-independent systems perform with an error rate that is about 3 or 4 times greater than similarly trained speaker-dependent systems (Pallett *et al.* [9]). Part of the problem is that speaker-independent systems like SPHINX have to cope with the burden of differing formant-frequency distributions from different speakers, which broaden the HMM distributions. In this section we present a novel technique for frequency normalization that is accomplished by multiplying the input cepstra by the matrix **L** in Eq. (1).

The frequency-normalization algorithm makes use of the bilinear transform stage already present in the SPHINX system, which accomplishes a nonlinear frequency warping of the cepstra. The bilinear transform is defined as

$$z_{new}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad -1 < \alpha < 1 \tag{13}$$

and it produces the frequency transformation

$$\omega_{new} = \omega + 2\,\mathrm{arctg}\left[\frac{\alpha \sin(\omega)}{1 - \alpha\cos(\omega)}\right] \tag{14}$$

In SPHINX we use the efficient algorithm proposed by Oppenheim and Johnson [10], that implements the bilinear transform as a matrix multiplication [7]. Specifically, the parameter $\alpha = 0.6$ is used to warp the LPC-cepstrum into a pseudo mel scale ([11]). The present algorithm selects a value of α to minimize the overall VQ distortion. This algorithm works in an *unsupervised* mode, since it does not require sex information or any other charac-

terization of the speaker's formant frequencies. The new codebook is generated by the Lloyd algorithm used for finding the VQ codebook, with the difference that the α parameter for every speaker is different.

In the present frequency-normalization algorithm, frequency warping is performed in two stages [7]: a bilinear transformation is first performed with $\alpha_0 = 0.6$, and then a second transform is applied with a variable warping parameter $\Delta\alpha$. The $\Delta\alpha$ for every speaker was chosen to minimize the VQ error while keeping the average $\Delta\alpha$ for 20 female speakers and 20 male speakers to be zero. Values of $\Delta\alpha$ examined ranged from -0.1 to 0.1 in increments of 0.02.

The use of frequency normalization increased the recognition rate from 85.3% to 87.1% using the DARPA standards for scoring (Pallett *et al.* [9]) and the census database. Re-running this algorithm on three independent test sets produced an average decrease in error rate of 10%. A histogram of the resulting warping parameters for the census database is shown in Figure 2 for male and female speakers. As we had anticipated there is a clear separation between them, which confirms our assumptions about this parameter.

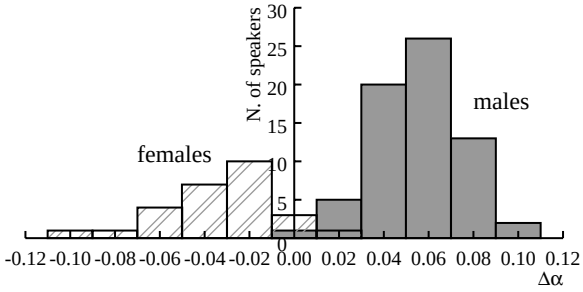

Figure 2: Histogram of values of α for male and female speakers estimated by the frequency normalization algorithm based on variable warping with the bilinear transform.

4. Conclusions

In this paper, we have presented several algorithms that normalize the acoustic space of the target speaker and environment, and that increase the robustness of the system to mismatches in training and testing conditions. An affine transformation is applied in the cepstrum vector with a frequency normalization matrix for different speakers and a correction vector for different acoustical environments. The correction vector and the warping parameter in the matrix are estimated to minimize the accumulated VQ distortion.

The ISDCN and FCDCN algorithms each make SPHINX more robust with respect to changes of microphone and acoustical environment. In ISDCN, the correction vector interpolates from the noise vector at low SNR to the equalization vector at high SNR by means of a sigmoid function. In FCDCN the correction vector is a function of the codeword chosen in the VQ stage as well as SNR. ISDCN is simple, efficient, and it does not require environment-specific training data. FCDCN is also computationally efficient and more accurate than ISDCN, but it does requires environment-specific data to derive the correction vectors. Finally, we introduced a novel method of frequency normalization that applies a different warping of the frequency axis to every speaker as to minimize the VQ distortion. This technique results in a 10% decrease in error rate compared to the baseline conditions.

5. Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 5167, under contract number N00039-85-C-0163. 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 Defense Advanced Research Projects Agency or the US Government. We thank Kai-Fu Lee, Robert Weide, Raj Reddy, and the rest of the speech group for their contributions to this work.

References

1. S. F. Boll, ''Suppression of Acoustic Noise in Speech Using Spectral Subtraction'', *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, April 1979, pp. 113-120.

2. D. Van Compernolle, ''Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings'', *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Albuquerque, NM*, April 1990, pp. 833-836.

3. T. G. Stockham, T. M. Cannon and R. B. Ingebretsen, ''Blind Deconvolution Through Digital Signal Processing'', *Proc. of the IEEE*, Vol. 63, No. 4, Apr. 1975, pp. 678-692.

4. A. Erell and M. Weintraub, ''Estimation Using Log-Spectral-Distance Criterion for Noise-Robust Speech Recognition'', *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Albuquerque, NM*, April 1990, pp. 853-856.

5. S. Seneff, ''A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing'', *Journal of Phonetics*, Vol. 16, Jan. 1988, pp. 55-76.

6. A. Acero and R. M. Stern, ''Environmental Robustness in Automatic Speech Recognition'', *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Albuquerque, NM*, April 1990, pp. 849-852.

7. A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, PhD dissertation, Carnegie Mellon University, Sep. 1990.

8. H. Gish, Y. Chow and J. R. Rohlicek, ''Probabilistic Vector Mapping of Noisy Speech Parameters for HMM Word Spotting'', *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Albuquerque, NM*, April 1990, pp. 117-120.

9. D. Pallett, J. G. Fiscus and J. S. Garofolo, ''DARPA Resource Management Benchmark Test Results, June 1990'', *Proc. Speech and Natural Language Workshop, Hidden Valley, PA*, Morgan Kaufmann, Jun. 1990.

10. A. V. Oppenheim and D. H. Johnson, ''Discrete Representation of Signals'', *Proc. of the IEEE*, No. 33, 1972, pp. 681-691.

11. Shikano, K, ''Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition'', Tech. report, Computer Science Department, Carnegie Mellon University, May 1986.