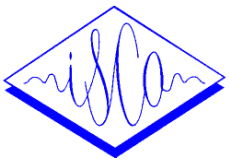

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

A UNIFIED APPROACH FOR ROBUST SPEECH RECOGNITION

Pedro J. Moreno, Bhiksha Raj, Richard M. Stern

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.

ABSTRACT

There are two major structural approaches to robust speech recognition. In the first approach to the problem, compensation is performed by modifying the incoming cepstral stream using ML or MMSE methods to estimate parameters characterizing environmental degradation, from direct frame-by-frame comparisons between speech recorded in high-quality and degraded acoustical environments, or by signal processing techniques such as spectral subtraction. The second approach tackles the problem by modifying the statistics of the internal representation of speech cepstra in the classifier to make them more closely resemble the statistics of degraded speech.

This paper attempts to unify these approaches to robust speech recognition by presenting three techniques that share the same basic assumptions and internal structure but differ in whether they modify the incoming speech cepstra or whether they modify the classifier statistics. We present SNR-dependent multivaRiate gAussian-based cepsTral normaliZation (SNR-RATZ) and SNR-based Blind RATZ (SNR-BRATZ), which modify incoming cepstra, along with STAR (STAtistical Re-estimation), which modifies the internal statistics of the classifier.

The algorithms were tested using the SPHINX-II speech recognition system on the CENSUS database, a database of strings of letters and numbers to which unknown added and unknown linear filtering was introduced artificially. While all the algorithms showed good performance, STAR was observed to provide lower error rates as SNR decreases than any of the algorithms that modify incoming cepstra.

1. INTRODUCTION

Speech recognition systems, like other pattern classification systems, consist of two parts: (1) signal processing that extracts meaningful features of the acoustic signal, which in our case are parametrized into cepstral coefficients, and (2) the classifier itself, which in our case is an HMM.

The effect of the environment on the speech recognition system is to alter the statistics of the incoming features. Compensation for environmental effects can be accomplished either using "signal processing approaches" that modify the cepstral parameters that are input to the system (*e.g.* [1,2]) or "classifier approaches" that modify the statistics of the representation of speech that is stored in the recognition system itself (*e.g.* [3]).

The standard signal processing approaches model the space of cepstra for clean speech either as a VQ codebook (*e.g.* [1,2]) or as a Gaussian mixture distribution (*e.g.* [4]) and they compensate for the changes in the parameters of this model due to the environment by applying an MMSE correction to the incoming noisy speech cepstra. The standard classifier adaptation approaches (*e.g.* [3]), on the other hand, adapt the statistics of the parameters characterizing speech as modeled by the HMM classification system to reflect the effect of the degradation on clean speech. Although some previous researchers have compared the performance of signal processing approaches and classifier adaptation approaches, these studies have either used different estimation structures to perform compensation for the two different approaches, or they used parameter estimation procedures that made use of less information in the signal than the algorithms described in this paper [5, 6, 7].

In this paper we present a unified approach to both methods of compensation. We base our approach on the knowledge of how the environment affects the statistics of speech cepstra and we obtain a common method of estimating these changes using a maximum-likelihood structure. Subsequently, we present three algorithms, SNR-dependent multivaRiate gAussian based cepsTral normaliZation (SNR-RATZ), SNR-based Blind RATZ (SNR-BRATZ), and STAtistical Re-estimation (STAR). SNR-RATZ and SNR-BRATZ are extensions of the RATZ and Blind-RATZ signal processing algorithms [4]: they modify the incoming speech cepstra to nullify the effects of the environment. STAR, on the other hand, is a classifier adaptation technique: it modifies the statistics of the HMMs that model the speech signal in order to adapt to the environment. SNR-RATZ and STAR make use of "stereo" adaptation data, which contains speech that is simultaneously recorded in the training and testing environment. SNR-BRATZ, on the other hand, makes use of "non-stereo" data, using speech samples that were recorded in the training and testing environments, but not necessarily simultaneously. All algorithms utilize the same ML structure.

In Section 2 we describe the three adaptation algorithms. In Section 3 we report the performance of each of these algorithms on the CENSUS database, an alphanumeric database consisting of sequences of strings and digits. Finally, in Section 4 we present our conclusions.

2. ALGORITHMS

Previous studies [4] have indicated that the effect of the environment on the statistics of speech can be well modeled as a shift of the means and reduction of the variance of the distribution of clean speech. Furthermore, when the distribution of clean speech is modeled as a combination of Gaussian distributions, as in the case of a Gaussian mixture distribution or a continuous or semi-continuous HMM, the effect of the environment is to shift the means and scale the variances of *each* of the Gaussians that comprise this distribution, where the changes in these parameters only depend on the original parameters of that particular Gaussian.

We model the distribution of the t^{th} vector, x_t of a cepstral vector sequence of length T, $\boldsymbol{X} = \{x_1, x_2, \ldots, x_T\}$, generically, as:

$$\sum_k a_k(t)\, N_k(\mu_k, \Sigma_k) \tag{1}$$

The overall likelihood for the observation sequence becomes

$$L(X) = \prod_{t=1}^{T} \sum_k a_k(t)\, N_k(\mu_k, \Sigma_k) \tag{2}$$

For the signal processing compensation methods, we set a_k to be independent of t; this defines a conventional Gaussian mixture distribution on the entire training set of cepstral vectors and is computed using standard EM methods. For the classifier adaptation methods, $a_k(t)$ represents the probability of being in state k at time t under the assumption that each state contains a single Gaussian[1]. These probabilities depend only on the Markov chain topology [8] and are represented in the usual form:

$$a(t) = \left[a_1(t)\ a_2(t) \ldots\ a_K(t)\right]^T = A^t \pi \tag{3}$$

where A represents the transition matrix and π the initial state probability vector of the HMM. The $N_k(\mu_k, \Sigma_k)$ s refer to the Gaussian densities associated with each of the states of the HMM.

The changes to the means and variances of the Gaussians can be expressed as:

$$\tilde{\mu}_k = \mu_k + r_k \qquad \tilde{\Sigma}_k = \Sigma_k + R_k \tag{4}$$

where μ_k and $\tilde{\mu}_k$ are the means of the k^{th} Gaussian in the distribution before and after contamination by the environment, r_k is the shift in the k^{th} mean, Σ_k and $\tilde{\Sigma}_k$ are the variances of the Gaussian before and after contamination by the environment and R_k is the change in the variance.

The generic ML estimates for the shift terms, r_k and R_k are obtained as:

$$\hat{r}_{x,k} = \left(\sum_{t=0}^{T-1} \gamma_t(k)\ (z_t - y_{tk})\right)\left(\sum_{t=0}^{T-1} \gamma_t(k)\right)^{-1} \tag{5}$$

$$\hat{R}_k = \frac{\sum_{t=0}^{T-1} \gamma_t(k)\ (z_t - \mu_k - \hat{r}_k)\ (z_t - \mu_k - \hat{r}_k)^T}{\sum_{t=0}^{T-1} \gamma_t(k)} - \Sigma_{x,k} \tag{6}$$

[1] An HMM with Gaussian mixtures for output PDFs can be reduced to the case where the output distributions are single Gaussians by considering the Gaussian mixtures as a set of Markov states with single Gaussians, where the incoming transition probability of the state is the same as the *a priori* probability of the Gaussians and the exiting transition probability is unity.

where z_t represents the vectors from the adaptation set and $\gamma_t(k)$ is the probability that the t^{th} observation of this set was generated by the k^{th} Gaussian. The definition of y_{tk} depends on whether or not the speech data used to estimate the parameters are stereo (*i.e.* simultaneously recorded in the training and testing environments). Specifically, y_{tk} represents the stereo counterpart of z_t; for non-stereo adaptation data $y_{tk} = \mu_k$. For the non-stereo case the equations become recursive and become the standard EM equations in the case of the signal-processing methods, and the Baum-Welch equations for learning HMM parameters in the case of the classifier adaptation method. The final step in the compensation process is an MMSE correction of the cepstral vectors in the case of the signal-processing methods and the correction of the HMM statistics in the case of the model adaptation methods.

2.1. SIGNAL PROCESSING ALGORITHMS: SNR-BRATZ AND SNR-RATZ

RATZ and Blind RATZ as described in [4] used a standard Gaussian mixture distribution as defined in Eq. (1) where the a_ks are taken to be independent of time to model the clean speech cepstral statistics. While this model of the clean speech cepstral statistics is good, it is constrained in that it resolves all the cepstral components equally, i.e into the same number of Gaussians. In particular, the frame energy parameter, x_0, has the same resolution in terms of number of Gaussians as the other cepstral parameters.

Previous work by Acero [1] and Liu [5] suggests that such fine modelling of x_0 may be unnecessary, and may in fact be detrimental to the performance of the algorithm as it is done at the cost of the modelling of the rest of the components in the cepstral vector.

SNR-RATZ and SNR-BRATZ use a more structured model for the distribution whereby the number of Gaussians used to define the x_0 statistics can be different from the number used for the other cepstral components. In these algorithms we use an implementation in which the zeroth component of $\boldsymbol{x}$, x_0, has a Gaussian mixture density. The statistics of the remaining components of $\boldsymbol{x}$, $\left[x_1\ x_2 \ldots\ x_P\right]$, are tied to the individual Gaussians that

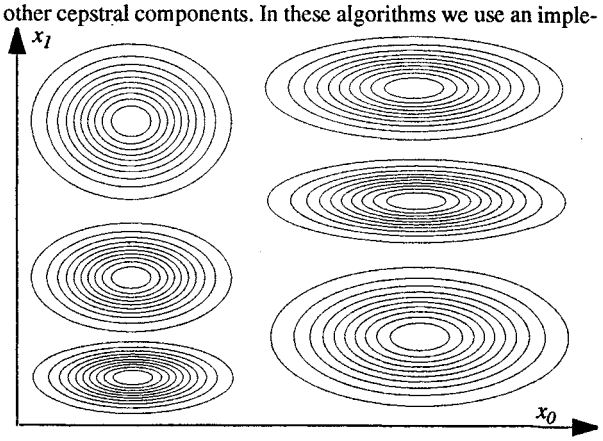

Figure 1. Contour plot illustrating joint PDFs of the structured Gaussian mixture densities for the components x_0 and x_1.

comprise the component x_0 to which they belong. Figure 1 illustrates the dependencies in Gaussian mixture densities that may arise for the first two cepstral components, x_0 and x_1. In this example the probability density function (pdf) for x_0 is a Gaussian mixture with two components. The pdf for x_1 that is associated with each component of x_0 is itself a mixture of Gaussians (with three components in this case). Note that the means of the mixtures that comprise the pdf of x_1 associated with each mixture component of x_0 can take on any value, and they generally differ for different values of x_0. The means, variances and *a priori* probabilities of the individual Gaussians are learned by standard EM methods.

The effect of the environment on the distribution is expressed as a shift of the means and a change of the variances of the Gaussians and are estimated using forms of Eq. (5) and Eq. (6). The exact form of the equation used is dependent on whether the adaptation data used to compute them is stereo or non-stereo.

Finally, noisy speech observations from the testing set are compensated by using an approximate MMSE technique to render them more like clean speech vectors:

$$\hat{x} = E(x|z) = \int_X x.p(x|z)\,dx = z - \int_X r(x)\,p(x|z)\,dx \equiv z - \sum_k P(k|z)\,r_k \quad (7)$$

2.2.1. SNR-BRATZ

SNR-based Blind RATZ (SNR-BRATZ) uses a non-stereo adaptation set of data that comes from the noisy environment to estimate the shifts in the means and changes of the variances of the Gaussians that comprise the distribution that models the clean speech statistics. These parameters are estimated using the Eq. (5) and Eq. (6) iteratively as follows:

$$\hat{r}_k^{l+1} = \left(\sum_i (z_i - \mu_k)\,P^l(k|x_i)\right)\left(\sum_i P^l(k|x_i)\right)^{-1} \quad (8)$$

$$R_k^{l+1} = \frac{\sum_i (z_i - \mu_k - \hat{r}_k^{l+1})(z_i - \mu_k - \hat{r}_k^{l+1})^T P^l(k|x_i)}{\sum_i P^l(k|x_i)} - \Sigma_k \quad (9)$$

where $P^l(k|x_i) = P(k|x_i, r^l_k, R^l_k)$

2.2.2. SNR-RATZ

When stereo data are available, this additional information can be utilized to obtain better estimates of the changes of the means and variances. We assume that the *a posteriori* probabilities of the Gaussians conditioned on the observed noisy frames, $P(k|z_i)$ do not change due to the environment and therefore can be well modeled by $P(k|x_i)$, the *a posteriori* probabilities conditioned on the corresponding clean speech frames. As the *a posteriori* probabilities are available, the estimation formulae for the changes in the means and variances are no longer iterative. These modified estimation formulae are

$$\hat{r}_k = \frac{\sum_i (z_i - x_i)\,P(k|x_i)}{\sum_i P(k|x_i)} \quad (10)$$

$$\hat{R}_k = \frac{\sum_i (z_i - \mu_k - \hat{r}_k)(z_i - \mu_k - \hat{r}_k)^T P(k|x_i)}{\sum_i P(k|x_i)} - \Sigma_k \quad (11)$$

2.2. CLASSIFIER ADAPTATION ALGORITHMS: STAR

STAR approaches the problem of robust speech recognition as a classifier adaptation problem which makes use of stereo adaptation data. The distribution for the speech cepstra are now HMMs as in Eq. (1) where the variables $a_k(t)$ follow a Markov chain.

The correction factors are computed as in Eq. (5) and Eq. (6) where y_{tk} is now the stereo counterpart of z_t. The variable $\gamma_t(k)$, which represents the probability of being in state k at time t, is conditioned on the statistics of the *clean* speech. Once again we assume that the *a posterior* probabilities $\gamma_t(k)$ do not change due to the effects of noise or filtering and can be computed from the clean speech.

This model is applied to the cepstra, delta-cepstra, and double delta-cepstra produced by SPHINX-II, along with a fourth three-dimensional stream that contains the cepstral component c_0, its difference Δc_0, and its double difference $\Delta^2 c_0$. In practice, we have observed that adapting the cepstral double-delta statistics does not affect the recognition performance. Once the correction terms are computed, the Gaussians are adapted to the new environment by changing their parameters as in Eq. (4).

3. EXPERIMENTAL RESULTS

Several experiments were performed to evaluate the word accuracy provided by SNR-RATZ, SNR-BRATZ, and STAR, along with selected related algorithms. The database used was part of the CENSUS database, and consists of strings of letters and digits [2], originally recorded with the close-talking Sennheiser HMD-414 microphone (CLSTLK). The training set consisted of 1014 sentences, recorded directly using the CLSTLK microphone. The test set contained 140 sentences. These sentences were first passed through a linear filter with transfer function shown in Fig. 2, and subjected to further corruption by stationary white gaussian noise at various SNRs.

In the first experiment we compared the recognition accuracies obtained by using the new compensation algorithms STAR and SNR-RATZ at each of the SNR levels. For SNR-RATZ, the distribution employed to model the clean speech made use of 4 Gaussians to model the x_0 component and 32 Gaussians associated with each of the x_0 Gaussians to model the rest of the cepstral vector. For comparison purpose, we also provide curves describing recognition accuracies obtained for three additional experimental conditions: (1) cepstral mean normalization (CMN) alone, (2) complete retraining using speech that had been degraded exactly in the same fashion as the test set, and (3) use of the cepstral compensation algorithm FCDCN [1],

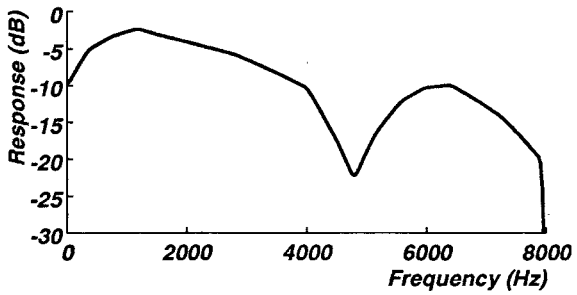

Figure 2. Frequency response of linear filter used to simulate channel effects.

which had been the most effective algorithm developed by our group before the present research began.

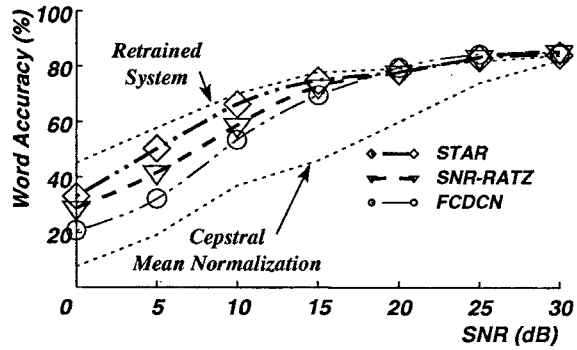

Figure 3. Comparison of recognition accuracy obtained using the signal processing algorithm SNR-RATZ, the classifier adaptation algorithm STAR, and our previous best algorithm, FCDCN. Testing speech had been corrupted by linear filtering and additive noise; results are plotted as a function of SNR.

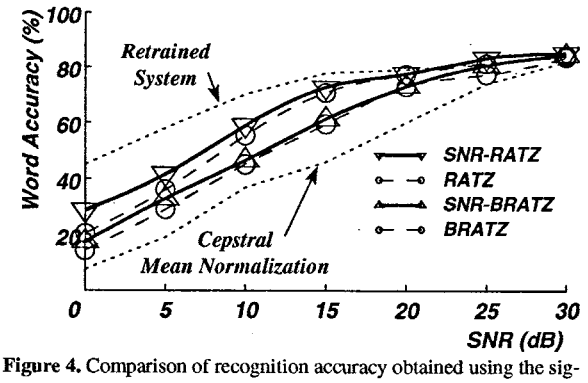

Figure 4. Comparison of recognition accuracy obtained using the signal processing algorithms SNR-RATZ, RATZ, SNR-BRATZ, and BRATZ.

We note that STAR outperforms all other algorithms and is only slightly worse than completely retraining the system to the noisy speech. We believe that STAR is superior, especially at low SNRs, because signal processing algorithms such as SNR-RATZ that attempt to correct for the effects of noise do not account completely for the changes of ideal classification boundaries that occur due to the effects of noise on the variances of the distributions. Furthermore, additional approximation errors are introduced in the MMSE process, leaving a residual mismatch between the estimates of "clean" speech and the original (clean) HMMs. In contrast, classifier adaptation algorithms such as STAR modify the variances as well as the means in the internal representation of the incoming features This is a better approximation to the ideal condition where training and testing are performed in the same environment.

Figure 4 compares the performance of SNR-RATZ and SNR-BRATZ with their simpler counterparts RATZ and BRATZ, implemented in the same fashion. We observe that SNR-RATZ and SNR-BRATZ perform slightly better than the non-SNR versions, particularly at lower SNRs. Unsurprisingly, the SNR-RATZ and RATZ, which are provided frame-by-frame comparisons of speech in the training and testing domains provide higher recognition accuracy than the "blind" algorithms (which do not require stereo data).

4. SUMMARY AND CONCLUSIONS

In this paper we present a unified approach to the compensation of speech parameters for the effects of the acoustical environment, either by modifying the incoming cepstral features or by adapting the classifier itself. We also present three algorithms as exemplars of the two approaches. While all the algorithms provided some benefit, the classifier adaptation algorithm was observed to produce the best recognition accuracy.

ACKNOWLEDGMENTS

We thank Evandro Gouvêa for running some of the experiments, as well as Uday Jain, and Alex Acero for useful suggestions. 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.

REFERENCES

1. Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.

2. Neumeyer, L., and Weintraub, M., "Probabilistic Optimun Filtering for Robust Speech Recognition", *Proc. ICASSP-94*, pp. 417-420, April, 1994.

3. Gales, M. J. F., and Young, S. J., "Cepstral Parameter Compensation for HMM Recognition in Noise", *Speech Communication*, **12**:231-239, 1993.

4. Moreno, P. J., Raj, B., Gouvêa, E. and Stern, R. M., "Multivariate Gaussian-Based Cepstral Normalization", *Proc. ICASSP-95*, pp. 137-140, May, 1995.

5. Liu, F-H., "Environmental Adaptation for Robust Speech Recognition". *Ph.D. Thesis*, ECE Department, CMU, July, 1994.

6. Sankar, A., and Lee, C.-H., "Robust Speech Recognition based on Stochastic Matching", *Proc. ICASSP-95*, pp. 121-124, May, 1995.

7. Legetter, C. J., and Woodland, P. C. (1994). "Speaker Adaptation of Continuous Density HMMs using Linear Regression". *Proc. ICSLP-94*, pp. 451-454, September, 1994.

8. Baum, L. E., Petrie, T., Soules, G. and Wiess, N., "A Maximumization Technique Orccuring in the Statistical Analysis of Probabilistic Functions of Markov Chains". *Ann. Math Statists.* **41**:164-171, 1970.