COVARIANCE ESTIMATION METHODS FOR CHANNEL ROBUST TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Michael Schmidt *Herbert Gish* *Angela Mielke*

BBN Systems and Technologies
70 Fawcett Street, Cambridge, MA 02138 USA

ABSTRACT

Two novel channel robust methods are described for performing text-independent speaker identification. The first technique models speaker's voices stochastically via cepstra correlations rather than by covariances in an effort to compensate for additive noise. The second technique, which we term dynamic covariances, models speakers by covariances of deviations of cepstra from time varying means rather than from constant means. Dynamic covariances may normalize for time varying channel effects, utterance lengths and text. Experimental results are obtained on the SPIDRE subset of the Switchboard corpus. Error rates as low as 2.2% are obtained using the new models.

1. Introduction

When test session channels differ from training channels, a variability in the data is introduced which makes the task of text-independent speaker identification more difficult. (By speaker identification we refer to the problem of chosing a speaker from a list of possible speakers as the speaker of a given utterance. When the text of an utterance is irrelevant to the identification procedure, the identification is termed text-independent.) Most popular techniques employ some form of mean removal to compensate for channel differences. One such method is to model speakers' voices by covariances of cepstra (Gish *et al.* [1,2]), thereby modeling by shape rather than by location. Additionally covariance matrices of cepstra are invariant to linear time-invariant channel effects. Some channel effects, for example additive noise, may, however, not be linear time-invariant. Even after mean removal, channel can be a significant source of error. In fact, on one experiment using the Switchboard corpus, we found that using covariance models, recognition accuracy decreased from 92% to 75% when test channels were present in training compared to when they were not.

Two covariance estimation techniques for coping with channel effects are presented. The first new algorithm is based on modeling speakers voices by correlations of cepstra rather than by covariances in order to compensate for varying noise levels. Additive noise reduces cepstral variance and so impacts on covariance matrices. The second technique, which is based on what we call dynamic covariances, attempts to compensate for time varying channel artifacts. Covariances are computed from deviations of cepstra from a time varying function. Dynamic covariances may also compensate for differences in the words spoken between training and testing, though we have found that dynamic covariances help little, if at all, when training and test channels are matched. Both methods are extensions of the robust segmental algorithms of Gish *et al.* [1,2] which are briefly reviewed below.

2. Background

The basic features used are mel-warped cepstra and difference cepstra. For each segment of speech, the sample mean $\overline{x}$, and sample covariance S of the cepstra as well as the sample covariance S_Δ of difference cepstra are computed. For each of several training sessions for each speaker, probability models for each of the above statistics are constructed. The distributions of the cepstra are assumed Gaussian in which case the mean statistic $\overline{x}$ is Gaussian and the covariances S and S_Δ have Wishart distributions. Let μ and Σ denote a model mean and covariance respectively estimated in training. Then the log likelihood of the mean of a test session $\overline{x}$ is

$$\ell(\overline{x}; \mu, \Sigma) = -\frac{1}{2}\log|2\pi n^{-1}\Sigma| - \frac{n}{2}(\overline{x}-\mu)'\Sigma^{-1}(\overline{x}-\mu), \quad (1)$$

and the log likelihood of S is

$$\ell(S; \Sigma) = -\frac{n-1}{2}\log|\Sigma/n| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S) + k_{S,n,p}. \quad (2)$$

($|A|$ and A' denote the determinant and transpose of a matrix A, respectively.) The formula for $\ell(S_\Delta; \Sigma_\Delta)$, the difference cepstra covariance log likelihood, is identical to Equation 2. The log likelihood of the mean of the difference cepstra depends only on the endpoints of an utterance, is random and is not useful. The log likelihoods of interest can be normalized by subtracting the mean and dividing by the standard deviation of the log likelihoods generated over all speaker models, weighted and combined to form new scores. Let MEAN, COV and DCOV denote the respective normalized log likelihoods. The weighted combined scores then take the form

$$\alpha\,\mathrm{MEAN} + \beta\,\mathrm{COV} + \gamma\,\mathrm{DCOV}. \quad (3)$$

The normalization enables the combining of disparate scores. The speaker associated with the model resulting in the maximum score is identified. When test session channels are different from training, it is usually beneficial to remove the mean feature term by setting $\alpha = 0$. Otherwise identification may be based on channel as much as on speaker.

0-7803-2431-5/95 $4.00

In order to combat contamination due to silence, noise or crosstalk, test sessions are segmented into 1 second intervals so that the best of multiple models from each speaker can be chosen for each segment and so that offending segment scores can be deemphasized or discarded. The segment scores are the log likelihoods of individual 1 second intervals and are combined using robust scoring algorithms. Three such algorithms are reviewed. The **Sum** score consists of each speaker's best segment scores summed. The **TopMtoN%**, is the sum of the top M% to N% speaker's ordered segment normalized scores. Segment scores are normalized by subtracting other speaker's scores from the same segment. Segment scores are normalized so that no one segment is weighted more than others and so that each speaker's scores will be combined from their individual best segments rather than all speakers scores generated from the same segments. The **Clip0** score is generated by summing each speaker's positive normalized segment scores. In effect, only segments where a speaker scores higher than all other speakers are counted in the score. Each robust score, for example Clip0, can be computed using means, covariances or derivative cepstra covariances.

Since the models are so simple, namely one or a few covariances per utterance, and since scoring is performed via the straightforward Equation 2, both training and testing are easily performed in real time using off-the-shelf workstations. No expensive searching, clustering or iterative algorithms are required.

3. Baseline Experiments

All experiments are performed using the SPeaker IDentification REsearch (SPIDRE) subset of the Switchboard corpus. The reader is referred to [3] for more information about Switchboard. The SPIDRE corpus consists of 45 speakers, 23 male and 22 female. Three 60 second training sessions were employed for each speaker. Results are based on one 40 second test per speaker. Each test contains approximately 30 seconds of speech from the speaker to be identified along with silence, noises and crosstalk. Handsets used in the test sessions were not encountered in training, hence the MEAN feature scores are not employed. Both COV and equally weighted COV+DCOV baseline results are presented in Table 1 for comparison with results in future sections. Note that error percentages are based on relatively few tests and that each incorrect identification results in a increase of 2.2% in the error rate.

Percent Error	COV	COV + DCOV
Sum	24.4%	11.1%
Clip0	24.4%	17.8%
Top0to60%	24.4%	13.3%
Top0to80%	24.4%	11.1%
Top10to80%	26.7%	8.9%

Table 1. Baseline covariance method results.

4. Correlation Models

A possible source of channel distortion is added noise. When noise is added to a signal, the energy of the signal goes up, but the magnitudes of the cepstral vectors tend to decrease [4,5]. Variances are reduced when magnitudes are decreased. To normalize for changes in cepstral variances due to added noise between training and testing, correlation matrices are substituted for covariance matrices.

Recall that the covariance for a collection of samples $\{x_i\}$ is given by the formula

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})'. \tag{4}$$

The correlation matrix R is obtained by dividing each entry s_{ij} of S by $s_i s_j$, where s_i is the standard deviation of the i^{th} coordinate. Equivalently, if D is the diagonal matrix with the same diagonal entries as the covariance matrix S, then

$$R = D^{-1/2} S D^{-1/2}. \tag{5}$$

The variances of the cepstral coefficients are normalized to one.

Two scoring algorithms using correlation matrices for identification are considered. The first technique assumes a Wishart distribution on correlation matrices and again uses Equation 2 replacing test and model covariances by test and model correlations, respectively. (The Wishart assumption is not unreasonable as the correlation of a collection of vectors is the covariance of the vectors scaled by their variances. Correlations can be thought of as covariances in a transformed space.) We denote the log likelihood by WISH-R for Wishart correlation. Slight improvements are obtained Experimentally. Compare Table 2 with Table 1.

Percent Error	WISH-R	WISH-R + DCOV
Sum	17.8%	11.1%
Clip0	17.8%	17.8%
Top0to60%	17.8%	6.7%
Top0to80%	17.8%	6.7%
Top10to80%	15.6%	8.9%

Table 2. Wishart density correlation results.

Alternatively, distributions of sample correlations in a rotated space can be used. The following fact is employed: The log density of a correlation matrix calculated using vectors sampled from a normal distribution with a diagonal covariance equals a constant times the log of the determinant of the correlation matrix plus another constant. [6, p. 266]. The model covariance matrices are routinely not diagonal, however a change in coordinates rectifies the situation. In the training stage, the eigenvectors E of the covariance matrix are computed in order to effect the desired change of coordinates. The covariance matrix in the new coordinate system, $E'\Sigma E$, is diagonal. The matrix

$$R^* = D^{-1/2}(E'SE)D^{-1/2}, \tag{6}$$

where D is now the diagonal matrix of $E'SE$, is the test correlation matrix in the rotated coordinate system. The log likehood of R^* is

$$\ell(R^*; E) = k_{n,p} \log|R^*| + k'_{n,p}. \tag{7}$$

We name $\ell(R^*; E)$, ROT-R.

Error percentages are reduced significantly by using the ROT-R scoring algorithm as illustrated in Table 3. When combining ROT-R and DCOV and the robust Top0to80% scoring algorithm is used, an error rate of 2.2% is obtained. Experimentally, correlation matrices possess a channel robustness which covariance matrices do not have.

Mansour and Juang [4] projected training cepstral vectors onto test vectors, resulting in scaled test vectors in order to improve

Percent Error	ROT-R	ROT-R + DCOV
Sum	13.3%	11.1%
Clip0	28.9%	15.6%
Top0to60%	20.0%	4.4%
Top0to80%	13.3%	2.2%
Top10to80%	15.6%	4.4%

Table 3. Rotated correlation density results.

speech recognition results when signal-to-noise ratios are low. They argue that cepstra directions are more robust to changes in the signal-to-noise ratio than cepstra magnitudes. In text-independent identification, however, individual test and training cepstra cannot be matched. Therefore, we experimented with scaling individual cepstral vectors to unit length. Recognition results, however, decreased dramatically. Even though cepstra magnitudes are susceptible to channel changes, they still provide useful identification information. Perhaps scaling by some function of the signal-to-noise ratio of utterances is the thing to do.

At first glance, replacing covariances by correlations appears to be liftering; namely a scaling the cepstral coefficients. Liftering, however, scales all sessions by a uniform amount, whereas we scale each training and test session separately. Liftering has no impact on WISH-R identification results. In fact, all linear transformations of the cepstral space affect only the inconsequential constant of Equation 2: If C is a linear transformation, then the sample covariance of the transformed cepstra, CSC', is again Wishart with parameter $C\Sigma C'$. Hence

$$\ell(CSC'; C\Sigma C') =$$

$$-\frac{n-1}{2}\log|C\Sigma C'/n| - \frac{n}{2}\mathrm{tr}(C'^{-1}\Sigma^{-1}C^{-1}CSC') + k_{S,n,p}$$

$$= -\frac{n-1}{2}\log|\Sigma/n| - (n-1)\log|C| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S) + k_{S,n,p}$$

$$= -\frac{n-1}{2}\log|\Sigma/n| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S) + k'_{S,C,n,p}, \tag{8}$$

since $C'^{-1}\Sigma^{-1}SC'$ has the same eigenvalues as $\Sigma^{-1}S$. ROT-R results are, however, not invariant to linear transformations and in particular liftering.

Correlations were also tested on the difference cepstra. After all, if the cepstra are scaled, then so are the difference cepstra. Experimentally, however, recognition results using difference cepstra correlations rather than difference cepstra covariances declined slightly.

5. Dynamic Covariances

In measuring a speaker's cepstral covariance matrix, we are measuring the covariance of an error in the fit of a model to data. In many cases the model is simply a mean vector and we are simply measuring the covariance of deviations of the observations to this relatively simple model. In general, however, a sequence of cepstral vectors can be represented as

$$x(t) = \mu_\theta(t) + \epsilon(t), \qquad t = 1, \ldots, n, \tag{9}$$

where $\mu_\theta(t)$ is the time varying mean with parameter θ, $\epsilon(t)$ is the deviation of the observation from the mean and n is the number of observations. With this model we estimate the covariance of the errors by

$$S = \frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}(t)\hat{\epsilon}'(t), \tag{10}$$

where

$$\hat{\epsilon}(t) = x(t) - \mu_{\hat{\theta}}(t) \tag{11}$$

and $\hat{\theta}$ is an estimate of the parameter of the mean $\mu_\theta(t)$. We refer to the covariances estimated from a model such as the one given by Equation 9 as Dynamic Covariances in that they are obtained from models having time varying means.

A motivation behind dynamic covariances is that by having a more accurate representation of the underlying process via the time-varying mean, the error term given by Equation 11 better characterizes the speaker. If the mean is not well estimated, then the variances of the error term will be inflated by the lack of fit. Speech is a non-stationary process with the sources of non-stationarity being the words spoken as well as time variations in the channel due to spurious noise events, changing noise levels, and hand-set movement, among others. Dynamic covariances are aimed at improving speaker identification performance in the face of these sources of variability.

Dynamic covariances are also applicable in those situations where the test speech is of significantly shorter duration than the amount of data employed in estimating the speaker's covariance matrix. For example, typically 30 seconds of training speech are used to obtain reasonable estimates for the 105 covariance parameters for the 14 low order cepstra. In contrast, test utterances are segmented into 1 second chunks for our robust scoring algorithms. Hence a mismatch in the standard covariances computed in testing and training. Variances are greater in training than in testing due to longer utterance lengths and hence a worse fit by a constant mean. Dynamic covariances can compensate for mismatches in covariances due to utterance length.

Table 4 gives dynamic covariance (DYN-COV) results when $\mu_{\hat{\theta}}(t)$ is piecewise constant over one second intervals. Since the means in training are computed over one second intervals the covariances in training now "match" the one second covariances computed in testing for the robust scoring algorithms. As test covariances are computed from exactly one second of speech normally, no changes are required in the test code. The DYN-COV results, alone are as good as the baseline COV+DCOV results, thereby eliminating the need to compute derivative features and scores altogether. Error rates obtained using piecewise linear $\mu_{\hat{\theta}}(t)$ dynamic covariances are slightly higher. In a similar vein, Montacie [7] and then Griffin *et al.* [8], use covariances of cepstra deviations from a Multivariate Auto-Regression (MAR) model, $M(t, x(t-1), \ldots, x(t-q))$, to perform speaker identification.

Percent Error	DYN-COV	DYN-COV + DCOV
Sum	13.3%	13.3%
Clip0	15.6%	17.8%
Top0to60%	11.1%	11.1%
Top0to80%	8.9%	11.1%
Top10to80%	8.9%	6.7%

Table 4. Piecewise constant $\mu(t)$ dynamic covariance results.

The dynamic covariance approach is not applicable for use with

difference cepstra as the difference cepstra means are all zero plus some negligible random error, as mentioned earlier. Combining techniques, a dynamic correlation approach is, however, possible. Table 5 gives results obtained combining the correlation and dynamic covariance methods. Piecewise constant mean correlation scores are computed and denoted by DYN-ROT-R. An error rate of 2.2% is obtained using dynamic correlations and the Top10to80% scoring algorithm.

Pct. Error	DYN-ROT-R	DYN-ROT-R + DCOV
Sum	15.6%	8.9%
Clip0	28.9%	15.6%
Top0to60%	13.3%	6.7%
Top0to80%	17.8%	4.4%
Top10to80%	17.8%	2.2%

Table 5. Piecewise constant $\mu(dynamic)$ correlation results.

6. Conclusion

Two new methods for text-independent speaker identification are presented. The first technique substitutes correlations rather than the more standard covariances to model speakers voices in an attempt to compensate for changes in cepstral variances due to changes in the signal-to-noise ratios between training and testing. The log likelihoods of correlations computed from Gaussian observations can be computed either by assuming they have Wishart distributions or by using the theoretical distribution of sample correlations in a rotated space. The second technique, dynamic covariances, models speakers by covariances of residuals from time varying mean models, $\mu_\theta(t)$, rather than residuals from constant mean model to help compensate for time-varying channel artifacts.

It is emphasized that too few tests were performed (and so too few errors are present) to make sweeping conclusions about the relative merits of the various algorithms. Nevertheless, the correlation scores, ROT-R, whether regular or dynamic, combined with DCOV scores result in the highest identification accuracies. Figure 1 presents a summary of the experimental results for the Top10to80% robust score. The hypothesis that the dynamic covariance score should improve results is reflected by the fact that the DYN-COV error rates drop from 26.7% to 8.9%. The decrease in error rate when the dynamic covariance is combined with DCOV is, however, less dramatic. Apparently, the information in DCOV is in some sense redundant with the information in the dynamic covariance. On the other hand, the ROT-R correlation results individually are not spectacular, but when ROT-R is combined with DCOV, error rates are greatly reduced.

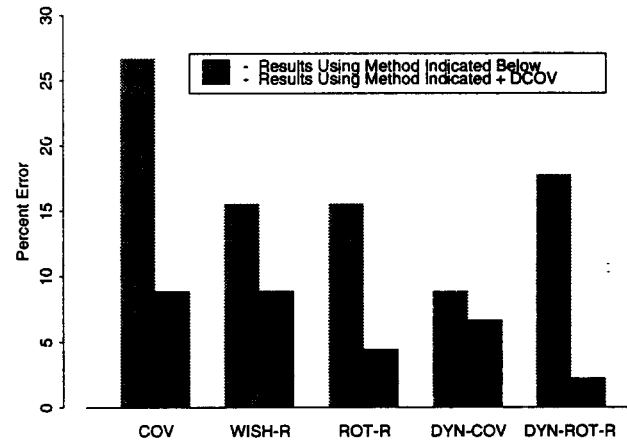

Figure 1. Summary of Top10to80% SPIDRE results

REFERENCES

[1] H. Gish, M. Schmidt, A. Mielke "A Robust Segmental Method for Text-Independent Speaker Identification", Proc. ICASSP '94, April 1994, Adelaide, South Australia, pp. 145-148.

[2] H. Gish, M. Schmidt, "Text-Independent Speaker Identification", IEEE Signal Processing Magazine, October 1994, pp. 18-32.

[3] J. Godfrey, E. Holliman, J. McDaniel, "Switchboard: Telephone Speech Corpus for Research and Development", Proc. ICASSP '92, March 1992, San Francisco, pp. 517 - 520.

[4] D. Mansour, B. H. Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition," Proc. ICASSP '88, April 1988, New York, pp. 36-39.

[5] L. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentiss Hall, New Jersey, 1993

[6] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis, 2nd Ed.*, J. Wiley & Son, N.Y., 1984.

[7] C. Montacie, "Cinematic Techniques for Speech Processing: Temporal Decomposition and Multivariate Linear Prediction," Proc. ICASSP '92, March 1992, San Francisco, pp. 153-156.

[8] C. Griffin, T. Matsui, S. Furui, "Distance Measures for Text-Independent Speaker Recognition Based on MAR Model," Proc. ICASSP '94, April 1994, Adelaide, South Australia, pp. 309-312.