Correspondence

Experimental Evaluation of Features for Robust Speaker Identification

Douglas A. Reynolds

Abstract— **This correspondence presents an experimental evaluation of different features and channel compensation techniques for robust speaker identification. The goal is to keep all processing and classification steps constant and to vary only the features and compensations used to allow a controlled comparison. A general, maximum-likelihood classifier based on Gaussian mixture densities is used as the classifier, and experiments are conducted on the King speech database, a conversational, telephone-speech database. The features examined are mel-frequency and linear-frequency filterbank cepstral coefficients, linear prediction cepstral coefficients, and perceptual linear prediction (PLP) cepstral coefficients. The channel compensation techniques examined are cepstral mean removal, RASTA processing, and a quadratic trend removal technique. It is shown for this database that performance differences between the basic features is small, and the major gains are due to the channel compensation techniques. The best "across-the-divide" recognition accuracy of 92% is obtained for both high-order LPC features and band-limited filterbank features.**

I. Introduction

The focus problem of the robust speech recognition workshop was to examine different ways to improve the robustness of speech recognizers, with particular emphasis on robust front-end speech analysis. However, in the course of the workshop, several related but different "underground" problems emerged; one was that of robust speaker identification. The task of speaker identification is closely related to speech recognition in that systems for both tasks generally use the same front-end speech analysis. Besides improving the robustness of speaker recognition systems, speech parameterizations providing robustness to speaker recognition systems might also be useful for speech recognition systems. However, since the underlying information in the speech waveform required by speaker-independent speech recognizers and text-independent speaker recognizers are very different, the robustness gains may not transfer between systems. Regardless, the potential application of speaker verification and identification for telephone-based transactions and speech-data information management merit research into robust speaker recognition in its own right.

This correspondence presents an experimental evaluation of different features for robust speaker identification. The speaker identification problem can be broadly divided into two components: speech analysis (or feature extraction) and classification. Although separate, these two components are tightly coupled. The feature space ultimately determines the separability of the desired classes (speakers), whereas the classifier must be correctly tuned to model and differentiate the classes in a given feature space. The emphasis of this paper is on the comparison of different feature spaces. Efforts were made in these experiments to keep all processing and classification steps constant and to vary only the features used. In this way, changes in performance can be attributed mostly to the features. A general, maximum-likelihood classifier based on Gaussian mixture densities, which has demonstrated excellent text-independent speaker recognition performance [1], [2], is used as the classifier. Experiments are conducted on the King speech database, which is a challenging and realistic conversational, telephone-speech, speaker recognition database.

The experiments compare four features and three channel compensation techniques used in different speech and speaker recognition systems reported throughout the literature. The features are mel-frequency and linear-frequency filterbank cepstral coefficients [3], linear prediction cepstral coefficients [4], and perceptual linear prediction (PLP) cepstral coefficients [5]. The channel compensation techniques are cepstral mean removal, RASTA processing [6], and a quadratic trend removal technique [7]. In comparing the different features, it is important to differentiate between performance gains due to the inherent robustness of the features and the robustness provided by the channel compensation techniques. It is shown that performance differences between the basic features is small and the major gains are due to the channel compensation techniques.

The rest of the correspondence is organized as follows. The next section briefly describes the Gaussian mixture speaker classifier and the King database used in the experiments. This is followed in Section III by a description of the features and channel compensation techniques compared. Section IV presents the experimental results. Finally, conclusions are given in Section V.

II. Classifier and Database

A. Gaussian Mixture Speaker Classifier

The speaker classifier models the distribution of features from a person's speech by a Gaussian mixture density. For a feature vector denoted as $\vec{x}_t$, the mixture density for speaker s is defined as

$$p(\vec{x}_t \mid \lambda_s) = \sum_{i=1}^{M} p_i^s \, b_i^s(\vec{x}_t). \qquad (1)$$

The density is a linear combination with weights p_i^s of M component Gaussian densities $b_i^s(\vec{x})$. Collectively, the parameters of a speaker's density model are denoted as λ_s. Given a sequence of feature vectors extracted from a person's training speech, maximum-likelihood parameter estimates are obtained using the EM algorithm.

In the experiments, each speaker is modeled by a 32-component GMM with nodal, diagonal covariance matrices. To avoid variance singularities, a variance floor of 0.001 was applied after each EM iteration, and nodal variances were lightly smoothed with the speaker's grand variance. Models were trained with 15 EM iterations initialized by a binary splitting k-means algorithm.

During recognition, a maximum-likelihood classification rule is used to identify the speaker of the input utterance. The log-likelihood score of a speaker's model for an input utterance feature vector sequence $\{\vec{x}_1, \ldots, \vec{x}_T\}$ is computed as $\mathcal{L}_s = \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_s)$. The speaker whose model produces the largest score is then determined to be the identified speaker.

Manuscript received December 1, 1993; revised April 22, 1994. This work was supported by the Department of the Air Force.

The author is with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02173 USA.

IEEE Log Number 9403966.

1063-6676/94$04.00 © 1994 IEEE

B. King Database

The experiments were conducted on the King database. The King database contains 10 short conversations from 51 adult male subjects recorded both locally using a high-quality electret microphone and after transduction by a carbon-button microphone and transmission over long-distance telephone lines. Except as noted, in the experiments described below, only the telephone speech was used. The recordings were made in two locations: Twenty-six speakers were recorded in San Diego, CA (SD) and twenty-five in Nutley, NJ (NJ). The first five sessions were recorded at nominal time intervals of one week and the second five sessions at intervals of one month. Each session is approximately 45 s in duration and contains at least 30 s of speech, excluding silence. The data is sampled at 8 kHz.

For the telephone speech, the NJ speakers' sessions are generally much noiser than the SD speakers' sessions. Using a peak signal-to-noise floor SNR estimate, the SD sessions had an average SNR of 30.8 dB, and the NJ sessions had an average SNR of 18.2 dB. Since sessions were transmitted over different long-distance lines, each session in general may have different channel filter shaping. However, the same carbon button handset was used for all recordings; therefore, there is no handset variability in the data.

For the SD speakers, some unknown change in the recording equipment or telephone lines between the recording of sessions five and six produced a large acoustic mismatch between the first five and second five recordings. This "divide" has been known for some time within the speaker recognition community and noted in a few papers [8]–[10]. The major effect is that good identification performance can be obtained if training and testing speech come from only one side of the divide, but performance drops dramatically if training and testing data come from different sides of the divide. Although this is possibly an anomaly of the database, it is an interesting challenge because the mismatch across the divide is not severe and is almost perceptually unnoticeable. Until this study, the best published performance on the 26 SD speakers is 94% within the divide and 78% going across the divide [10].

Speech/silence discrimination was done using an adaptive, energy-based thresholding algorithm [2]. For each session, a speech label file was created that marked 20 ms speech frames every 10 ms as speech or silence. These speech/silence markings are used in all the experiments so that the identical speech frames are used in training and testing for all features. The SD sessions had an average duration of 20 s after speech detection, whereas the noisier NJ sessions had an average of 12 s.

The King database was split into two speaker populations and four train/test divisions for the experiments. Due to the widely differing noise levels between the SD and NJ sessions, the SD and NJ speakers were used as different populations. Results are shown on the 26 SD speakers (denoted SD26) and the 25 NJ speakers (denoted NJ25) separately. The training and testing sessions were selected to examine within the divide (WD) and across the divide (AD) performance. Even though there is no divide in the NJ sessions, they were split the same way for consistency of presentation. The train/test sessions splits came out as follows:

WD: Train on sessions (1, 2, 3) test on sessions (4, 5)
Train on sessions (6, 7, 8) test on sessions (9, 10)
AD: Train on sessions (1, 2, 3) test on sessions (9, 10)
Train on sessions (6, 7, 8) test on sessions (4, 5)

All speech in a session is used to make an identification decision for testing. Reported WD results are the percentage of correctly identified sessions over both the WD train/test divisions. Likewise for the AD results. There were 104(=26 × 4) test sessions in both the WD and AD splits for the SD26 population and 100(=25 × 4) test sessions for the the NJ25 population. This gives a worst-case ± 5% binomial significance interval about the reported results.

III. Features and Channel Compensation

A. Features

The four feature sets used in the experiments are mel-frequency spaced filterbank cepstral coefficients (MFCC) [3], linear-frequency spaced filterbank cepstral coefficients (LFCC),[1] linear prediction cepstral coefficients (LPCC) [4], and perceptual linear prediction cepstral coefficients (PLPC) [5].

All the features examined are based on spectral information derived from a short time-windowed segment of speech. They differ mainly in the detail of the power spectrum representation. The filterbank features (MFCC and LFCC) are derived directly from the FFT power spectrum, whereas the linear prediction-based features (PLPC and LPCC) use an all-pole model to represent the smoothed spectrum. The mel-scale filterbank centers and bandwidths are fixed to follow the mel-frequency scale, giving more detail to the low frequencies, whereas the linear filterbank provides equal detail to all frequencies. The LPCC can be considered as having adaptive detail in that the model poles move to fit the spectral peaks wherever they occur. The detail is limited mostly by the number of poles available. The PLPC features are a hybrid between the filterbank and all-pole model spectral representation. The spectrum is first passed through a bark-spaced trapezoidal-shaped filterbank and then fit with an all-pole model. The detail of the PLPC representation is determined by both the filterbank and the all-pole model order.

The spectral representations are transformed to cepstral coefficients as a final step. This is done because of the (near) orthogonalizing property of the cepstral transformation. The filterbank representations are transformed directly by a discrete cosine transform (DCT). The all-pole representations are transformed using the recursive formula between prediction coefficients and cepstral coefficients [4]. In all cases, the zeroth cepstral coefficient is discarded as a form of energy normalization.

B. Channel Compensation

Three different forms of channel compensation are used in conjunction with the above features. These compensation techniques are all based on a linear filter model of the channel and rely only on the received speech signal.

Cepstral Mean Removal [4], [11]: In the cepstral domain, the filter response will appear as an additive component on each cepstral feature vector. In this technique, the cepstral average over the speech frames in a session is subtracted from each cepstral feature vector prior to training or recognition. Removing the global cepstral mean also removes the average speech spectrum, which can improve robustness to intersession variabilities [1].

RASTA Processing [6]: This technique is based on the same idea as the cepstral mean removal but focuses on short-term mean removal appropriate for rapidly time-varying channels. Each frequency component or filterbank output is passed through an ARMA filter that performs frame differencing (MA portion) and exponentially decaying mean subtraction (AR portion).

Quadratic Trend Removal [7]: This technique is based on the assumption that the channel filter is relatively smooth across frequencies compared with the speech spectrum and can be modeled

[1] Same as mel filterbank, but filter centers are equally spaced, and bandwidths are constant over the signal bandwidth.

TABLE I
BASELINE SPEAKER ID RESULTS IN PERCENT CORRECT. WD = WITHIN DIVIDE, AD = ACROSS DIVIDE, SD26 = 26 SAN DIEGO SPEAKERS, NJ25 = 25 NEW JERSEY SPEAKERS.

	SD26		NJ25	
Feature	WD	AD	WD	AD
MFCC-24 no chan. comp	75	15	40	39
MFCC-24	100	86	63	66
LFCC-40	100	84	63	64
PLPC-12	95	74	65	55
LPCC-12	99	80	62	59

by a quadratic polynomial in the log spectrum domain. The least-mean-squares (LMS) fit quadratic polynomial is subtracted from each log-spectrum speech frame prior to cepstral transformation to remove the channel filter effects. Note that this technique requires only one speech frame to estimate and remove the channel filter response.

The quadratic fit is best done over the passband of the channel filter (e.g., 400–3200 Hz) to avoid fitting nonessential transition and stop-band channel filter characteristics. This is done by performing the quadratic trend removal only over the filter outputs that fall in the channel filter's passband.

IV. RESULTS

This section presents the experimental results on the King database using the above features and channel compensation techniques. First, baseline results are presented for the basic feature sets. Next, recognition performance of the features for varying model orders (i.e., number of filterbanks, LPC order) is given. Finally, results comparing the different channel compensation techniques are presented.

A. Baseline

Table I gives the baseline performance of the different features. Baseline is defined as using cepstral mean removal and the following specifications for each feature set: MFCC-24 (24 mel-frequency spaced filters, 23 cep coeff/vector), LFCC-40 (40 linear-frequency spaced filters, 39 cep coeff/vector), PLPC-12 (17 Bark-frequency spaced filters, 12th-order LPC, 12 cep coeff/vector), LPCC-12 (12th-order LPC, 12 cep coeff/vector).

The performance for MFCC without cepstral mean removal is also given to show the need for channel compensation on this database (the other features also performed as poorly without some form of channel compensation).

It is clear that without some form of channel compensation, the performance is very poor. Just using simple cepstral mean removal dramatically improves performance, giving a 25 percentage-point increase for WD and a 71 point increase for AD for the SD26 speakers. However, there is still a 13-point gap between the WD and AD performance. The NJ25 speakers also have an increase of 25 percentage points but, due to the high noise levels, are performing much worse than the SD26 speakers.

Referring to the SD26 subset, all features tend to perform very well on the WD data, with PLPC performing slightly worse than the others. However, for the AD data, the PLPC and LPCC features perform worse than MFCC and LFCC features.

B. Model Orders

As mentioned earlier, the main differences between all the features examined in the level of detail each uses in representing the power spectrum will be described. The results in Table II compare the speaker ID performance for the different features using various "model" orders (the baseline results are also included in the table for comparison). Model orders means the number of filters in LFCC or the all-pole filter order in PLPC and LPCC. The MFCC model order is fixed at 24 mel-spaced filters. The cepstral order of the feature vector is also increased to match the model order, except as noted. Again, cepstral mean removal was used for channel compensation. The features and their model orders are MFCC-24c12 (24 mel-frequency spaced filters, 12 cep coeff/vector), LFCC-64 (64 linear-frequency spaced filters, 63 cep coeff/vector), PLPC-23 (17 Bark-frequency spaced filters, 23rd-order LPC, 23 cep coeff/vector), LPCC-23 (23rd-order LPC, 23 cep coeff/vector), LPCC-23c12 (23rd-order LPC, 12 cep coeff/vector), LPCC-46 (46th-order LPC, 46 cep coeff/vector).

TABLE II
SPEAKER ID RESULTS FOR VARYING FEATURE MODEL ORDERS IN PERCENT CORRECT. WD = WITHIN DIVIDE, AD = ACROSS DIVIDE, SD26 = 26 SAN DIEGO SPEAKERS, NJ25 = 25 NEW JERSEY SPEAKERS.

	SD26		NJ25	
Feature	WD	AD	WD	AD
MFCC-24	100	87	63	66
MFCC-24c12	99	86	59	55
LFCC-40	100	84	63	64
LFCC-64	98	83	59	65
PLPC-12	95	74	65	55
PLPC-23	96	75	65	65
LPCC-12	99	80	62	59
LPCC-23	100	* 92 *	62	61
LPCC-46	99	86	62	64
LPCC-23c12	98	81	57	59

The most surprising result is that a high-order LPCC feature set (order = 23) gave a 92% identification accuracy on the SD26 AD data. This is surprising for several reasons. First, speaker and speech recognition systems generally use LPC of orders 10–12 and it assumed that using much higher order models will model superfluous information or produce spurious peaks and thus degrade performance. Second, it is often stated that LPC models work well in clean environments but significantly degrade under harsher telephone environments. Even with the noisy NJ25 speakers, the LPC-23 and LPC-46 did not degrade worse than any other feature. Note that other researchers have also reported good speaker recognition results using high-order LPC [12].

Increasing the model order helped for the PLPC and LPCC features but slightly degraded the LFCC features. This probably occurs because the LPC model-based features can use the extra degrees of freedom to model spectral detail wherever it occurs in the frequency band, whereas the linear filter features increase spectral detail uniformly over the entire band, giving detail to superfluous information such as stop-band regions. It is important that the increased spectral detail not be used to model nonspeaker specific information, such as the channel or noise.

In the cases where the number of cepstral coefficients was less than the model order (MFCC-24c12 and LPCC-23c12), performance decreases, most significantly for the noisy NJ25 speakers. Using a reduced number of cepstral coefficients is equivalent to low-pass filtering the log spectrum, which further reduces the detail of the spectrum being modeled. For very noisy speech, this has the added

TABLE III
SPEAKER ID RESULTS COMPARING CEPSTRAL MEAN REMOVAL TO RASTA PROCESSING IN PERCENT CORRECT. WD = WITHIN DIVIDE, AD = ACROSS DIVIDE, SD26 = 26 SAN DIEGO SPEAKERS, NJ25 = 25 NEW JERSEY SPEAKERS.

	SD26		NJ25	
Feature	WD	AD	WD	AD
PLPC-12 cep mean removal	95	74	65	55
PLPC-12 RASTA	86	52	57	53

TABLE IV
SPEAKER ID RESULTS COMPARING CEPSTRAL MEAN REMOVAL, BANDLIMITING AND QUADRATIC TREND REMOVAL IN PERCENT CORRECT. WD = WITHIN DIVIDE, AD = ACROSS DIVIDE, SD26 = 26 SAN DIEGO SPEAKERS, NJ25 = 25 NEW JERSEY SPEAKERS.

	SD26		NJ25	
Feature	WD	AD	WD	AD
MFCC-24 cep mean removal	100	87	63	66
MFCC-24 bandlimiting (400-3200Hz)	97	28	50	51
MFCC-24 cep mean removal +bandlimiting (400-3200Hz)	96	91	63	67
MFCC-24 quad trend removal +bandlimiting (400-3200Hz)	93	78	58	57
LFCC-40 cep mean removal	100	84	63	64
LFCC-40 cep mean removal +bandlimiting (400-3200Hz)	98	92	64	64
LFCC-40 quad trend removal +bandlimiting (400-3200Hz)	93	81	58	59

effect of smearing noise-corrupted regions of the spectrum with uncorrupted regions.

C. Channel Compensation

In this section, results are presented that compare the channel compensations described in Section III-B to the baseline cepstral mean removal. Not all compensation techniques are suitable for all features, and not all possible combinations were tried to avoid a combinatorial explosion in experiments.

1) PLP-RASTA: It was found that RASTA processing improves performance over no compensation but does worse than cepstral mean removal (see Table III). Although RASTA processing does help remove channel effects, it appears that the adaptive mean removal operates over too short of a time span to compute a stable channel mean estimate. It is also possible that the frame differencing component of the RASTA filter is hurting the overall performance since it is known that using delta cepstral coefficients alone performs worse than static cepstral coefficients [13].

2) Quadratic Trend Removal: There are actually two effects examined here. First, since the quadratic trend removal is best suited for modeling the channel filter over the passband, the effect of using a band-limited spectrum is examined. Second, the effect of removing the quadratic trend is examined.

Results are presented for the MFCC and LFCC features in Table IV. The bandlimiting is done by performing cepstral analysis only over filterbank outputs that correspond to the channel filter's passband–empirically determined to be 400–3200 Hz. The lower frequency cutoff of 400 Hz is higher than the normally used 300-Hz cutoff. This value gave better performance probably because the 400-Hz cutoff is fully in the passband, whereas 300 Hz is actually in the transition region of the channel filter.

First, note that bandlimiting alone (no channel compensation) brings the WD performance almost to 100% and slightly boosts the AD performance, although it is still poor. This indicates that the major degradation found in the WD data is not spectral shaping mismatch but bandlimiting effects. Combining bandlimiting with cepstral mean removal then brings the AD performance to 91% for MFCC and 92% for LFCC. These values match the LPCC-23 results given earlier.

The quadratic trend removal also does a remarkable job of increasing the AD performance. AD performance increases 50 percentage points, whereas the WD performance drops about three points. Combining cepstral mean removal followed by quadratic trend removal improved performance over quadratic trend removal alone but was lower than cepstral mean removal alone.

V. CONCLUSION

Several observations can be made from these experiments. First, no feature set was inherently immune to the channel degradations. All features performed poorly on the AD data without some form of channel compensation. Second, increased spectral resolution beyond that commonly used for speech recognition helped performance for all features. Using only low-order filterbank cepstral coefficients hurt speaker ID performance compared with using all available cepstral parameters. Increasing the LPC order to a rather high order of 23 gave the best AD performance on King (92%). Of course, giving too much spectral resolution will degrade performance by modeling spurious spectral events or introducing too many parameters to be trained (LPCC-46, LFCC-64). With appropriate bandlimiting, the LFCC-40 and MFCC-24 features produced 92% and 91% accuracy on the AD data, respectively. Third, simple cepstral mean removal was the best channel compensation technique for all features. RASTA processing boosted performance over no compensation but performed worse than cepstral mean removal. Quadratic trend removal coupled with bandlimiting provided a large boost in performance for the AD data and was based on only a single frame data. Fourth, the filterbank spacing does not appear to be a critical factor for speaker ID. The linear and mel-scale filterbanks performed about the same in these experiments. However, there should be enough filters to model the speech bandwidth with sufficient detail. The comparatively lower performance of the PLPC features may be due to having only 17 bark-spaced filters to model the spectrum.

These experiments also give some insight into the artifacts found in the King database that degrade performance. Bandlimiting seems to be the major degradation in the SD26 subset WD data. Without any channel normalization, simply performing cepstral analysis over the passband (400–3200 Hz) brought the WD data performance to almost 100%. Since WD performance is around 100% for all features extracted over the full band using cepstral mean removal, it appears that cepstral mean removal helps eliminate most out-of-band artifacts. The high-order LPC worked without bandlimiting because it had more degrees of freedom (poles) to model the high-amplitude spectral region of the passband. Low-order LPC also places its resolution in the pass-band region, but a disproportionate amount of poles are allocated to model out-of-band regions.

A change in the channel filter spectral shape coupled with the bandlimiting appears to be the major effect causing such poor performance on the SD26 subset AD data. Visual inspection of the long-term power spectra from sessions on both sides of the divide clearly shows a change of spectral shaping. Although other effects such as noise and nonlinear distortion can be found in the data, the fact that cepstral mean removal coupled with bandlimiting almost equalizes AD and WD performance points to a predominant linear filtering effect.

For the NJ25 subset, the overriding degradation is very high noise levels. Although no divide exists in the NJ25 subset, cepstral mean removal improves performance for both AD and WD data, indicating that channel effects are also present. Further performance increases in the NJ25 subset will require noise compensation techniques.

In general, for the features examined in this investigation, the base features used appear less important than the channel compensation techniques applied to them. A majority of the features proposed for speaker recognition systems are based on some modeling of the power spectrum. If the spectral model does not overly smooth spectral details or provide too much detail to spurious spectral events, then there should be little difference in performance of the features. The main difference in features is how severely the features are affected by channel and noise effects and how easily these effects can be decoupled from the underlying speech spectral information. Unless a very unique speaker-dependent feature set is discovered, the overall performance of a speaker recognition system will be most heavily tied to the effectiveness of the channel compensation techniques and the classifier employed.

REFERENCES

[1] R. C. Rose and D. A. Reynolds, "Text-independent speaker identification using automatic acoustic segmentation," in *Proc. ICASSP-90*, 1990, pp. 293–296.

[2] D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification," Ph.D. Thesis, Georgia Inst. Technol., 1992.

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, Aug. 1980.

[4] B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.

[5] H. Hermansky, "Perceptual linear prediction (plp) analysis for speech," *JASA*, pp. 1738–1752, 1990.

[6] H. Hermansky *et al.*, "RASTA-PLP speech analysis technique," in *Proc. ICASSP-92*, Mar. 1992, pp. I.121–I.124.

[7] B. Mistretta, D. Morgan, and L. Rieck, "Experiments with open set speaker identification," Tech. Rep. VCI-5, Sanders, Dec. 1990.

[8] H. Gish, "Robust discrimination in automatic speaker identification," in *Proc. ICASSP-90*, Apr. 1990, pp. 289–292.

[9] A. Higgins, L. Bahler, and J. Porter, "Voice identification using nearest-neighbor distance measure," in *Proc. ICASSP-93, 1993*, pp. II-375–II-378.

[10] Y. Kao, J. Baras, and P. Rajasekaran, "Robustness study of free-text speaker identification and verification," in *Proc. ICASSP-93*, 1993, pp. II-379–II-382.

[11] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.

[12] G. Velius, "Variants of cepstrum based speaker identity verification," in *Proc. ICASSP-88*, 1988, pp. 583–586.

[13] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Proc. ICASSP-86*, 1986, pp. 877–880.