Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch

Sarel van Vuuren

Oregon Graduate Institute of Science & Technology
Center for Spoken Language Understanding
20000 N.W. Walker Road
Beaverton, Oregon 97006 USA

ABSTRACT

We compare speaker recognition performance of Vector Quantization (VQ), Gaussian Mixture Modeling (GMM) and the Arithmetic Harmonic Sphericity measure (AHS) in adverse telephone speech conditions. The aim is to address the question: how do multimodal VQ and GMM typically compare to the simpler unimodal AHS for matched and mismatched training and testing environments. We study identification (closed set) and verification errors on a new multi-environment database. We consider LPC and PLP features as well as their RASTA derivatives. We conclude that RASTA processing can remove redundancies from the features. We affirm that even when we use channel and noise compensation schemes speaker recognition errors remain high when there is acoustic mismatch.

1. INTRODUCTION

In a realistic telephone application, speech collected during enrollment of the speaker and available for initial training typically come from a single environment, while at test time the environment is generally unknown. Reynolds[10] observed that acoustic mismatch due to different training and testing environments can severely degrade recognition performance.

In this paper we study recognition performance for speech collected over different telephone handsets and channels. We apply LPC and PLP analysis and channel and noise compensation by RASTA [4], Cepstral Mean Subtraction (CMS) [3] and normalization [6]. We use Vector Quantization (VQ) [7], Gaussian Mixture Modeling (GMM) [10] and the Arithmetic Harmonic Sphericity measure (AHS) [1].

While results for mixture modeling are available on TIMIT, NTIMIT, Switchboard, YOHO and King databases [10] and results for unimodal statistical methods are available on Switchboard [11] and TIMIT and NTIMIT [2] these databases generally allow only limited cross-environment experiments (although some experiments with Switchboard[1] and King are possible [9]).

In contrast, the database used in this paper[2] allows us to explicitly investigate the effect of well characterized environments. This database consists of isolated words spoken by 36 speakers[3] from each of 4 different telephone handset and channel environments. Each speaker produced 6 repetitions of a fixed 13 word vocabulary in each environment. To relate results on this multi-environment database to results reported in the literature[2] we report[4] 35.7% (PLP) and 29.1% (LPC) identification error when using GMM over the 168 speakers in the test portion of the NTIMIT database, where for each speaker, we tested individually on two of the sx sentences and trained on the eight other sentences.

In the rest of this paper, we study the effect of different training and test environments on identification and verification performance.

2. DESCRIPTION

2.1. Features

PLP and LPC analysis. We compare the discriminability and robustness to noise of Perceptual Linear Prediction (PLP) [4] and Linear Prediction (LPC). For PLP the spectral scale is the non-linear Bark scale and the spectral features are smoothed within frequency bands. In contrast for LPC the spectral scale is linear and no smoothing is done.

For results reported in this paper speech sampled at 8Khz is analyzed within a 20ms window at a 10ms frame rate. After DC removal and speech/non speech detection the analysis steps for LPC are pre-emphasis (0.90) and all-pole modeling. The analysis steps for PLP are critical band warping and averaging (within 35 bands spaced according to the Bark scale), equal loudness pre-emphasis, transformation according to the intensity loudness power law and all-pole modeling.

[1] In Switchboard for example it has to be assumed that telephone numbers identify unique handsets.

[2] To be made available `http://www.cse.ogi.edu/CSLU/`

[3] The calls from this set of speakers have been verified manually.

[4] These values are for the same system as used on the multi-environment database – we did not explicitly try to optimize performance for the NTIMIT database.

All-pole model parameters are converted to cepstral coefficients which are liftered (1.0) to approximately whiten the features. For a model of order p we use $p = 20$[5] cepstral coefficients exclusive of the energy coefficient[6] to form feature vectors x_t at the frame rate. We do not use feature vectors with energy smaller than an adaptive threshold — assuming that these vectors may be non-speech or noisy.

Noise compensation and channel equalization. With speech coming from an unknown, likely noisy environment, as with the different channels and handsets used for the experiments in this paper, noise compensation and channel equalization might improve robustness. We compensate for convolutional noise (as may be due to the additive effect of a channel in the log-spectral domain), by subtracting the long-term average from the ceptral coefficients (CMS) on a per utterance basis and/or by bandpass filtering in the log-spectral domain (RASTA filtering [5]). CMS performs a lowpass filtering operation of the speech in contrast to the bandpass filtering (between 1Hz and 16Hz modulation frequency [5]) performed by RASTA.

In [6] it is shown that the norm of the cepstral coefficient vector is particularly sensitive to additive noise, while its direction is less affected. However, we found no benefit from normalizing the cepstral feature vectors to unit magnitude above and beyond the results reported here.

2.2. Models

It is of interest to compare recognition errors for the different modeling methods with respect to their associated representational power and robustness given noisy conditions. The motivation for considering multimodal (VQ, GMM) vs unimodal models (AHS) is that the multimodal models can model non-linear correlations (caused for example by the presence of different linguistic units in the speech) whereas the unimodal models are restricted to modeling linear correlations. Given suitable regularization and relatively clean speech, greater modeling accuracy is expected from multimodal models than their unimodal counterparts. Conversely unimodal models are expected to be less sensitive to small perturbations in the speech that might arise in noisy conditions. Of the modeling methods GMM has the greatest degree of modeling freedom.

Vector Quantization. We use VQ [7] as the baseline for results reported here. It differs from GMM in that it does not use local covariance information and classification takes place in a winner-takes-all fashion. We train the VQ models with the LBG algorithm, choose to model the feature vectors for each speaker by 32 clusters and use the Euclidean norm.

Gaussian Mixture Modeling. GMM [10] uses a mixture of Gaussian densities to model the distribution of the feature vectors x of each speaker. For M mixtures the mixture density for speaker r is modeled as $p(x|\theta_r) = \sum_{i=1}^{M} \alpha_i^r p(x|i, r, \mu_i^r, \Sigma_i^r)$, with the restrictions $\alpha_i^r > 0$ and $\sum_{i=1}^{M} \alpha_i^r = 1$. The $p(x|i, r, \mu_i^r, \Sigma_i^r)$ are multivariate normal densities with mean vector μ_i^r and covariance matrix Σ_i^r. Here θ_r denotes the parameter vector $(\alpha_i^r, \mu_i^r, \Sigma_i^r)_{i=1}^n$. We estimate the parameter vector with the EM algorithm and regularize with a Bayesian prior [8]. Since AHS covers the case of a full single covariance matrix effectively, we choose to use a 32 mixture GMM with diagonal covariance matrices [10].

Given a reference model θ_r for speaker r and assuming independent feature vectors $X = \{x_1, \ldots, x_T\}$ the average log-likelihood for the utterance is formulated as $\mathcal{L}(X|\theta_r) = 1/T \sum_{t=1}^{T} \log p(x_t|\theta_r)$. Assuming that local covariance information is preserved in adverse conditions and can be accurately estimated, we expect GMM to outperform VQ on average.

Arithmetic Harmonic Sphericity measure. AHS [1] is a function of the eigenvalues of a test covariance matrix S relative to a reference covariance matrix S_r for speaker r and is defined by the measure $d_r = \log\left[\mathrm{tr}(S_r S^{-1})\mathrm{tr}(S S_r^{-1})\right] - 2\log D$, which is non-negative and zero iff all the eigenvalues are equal.

2.3. Tasks

Identification. The closed-set identification task which we consider here is to classify speech from data X as belonging to speaker $\hat{r}$ for which

$$\hat{r} = \arg\max_r \Pr(\theta_r|X) \propto \arg\max_r \mathcal{L}(X|\theta_r) \qquad (1)$$

using Bayes' rule and assuming equal prior probabilities of speakers. When the likelihood is viewed as a distance the task is $\hat{r} \sim \arg\min_r d_r$.

Verification. Given a claimant speaker $\hat{r}$ and data X the verification task which we consider here is the hypothesis test $H_0 : X$ is from $\hat{r}$, vs $H_1 : X$ is not from $\hat{r}$, where the decision is taken reject H_0 iff $\lambda < \lambda_c$ for a log likelihood ratio

$$\lambda = \log(\Pr(\theta_{\hat{r}}|X)) - \log(\Pr(\theta_{r\neq\hat{r}}|X)). \qquad (2)$$

By taking the θ_r from a set of representative imposters or *cohorts* the effect of the second term in Eq. 2 is to normalize the likelihood for the data from speaker $\hat{r}$. For GMM we apply Bayes' rule as in Eq. 1 and approximate $\lambda = \mathcal{L}(X|\theta_r) - \log\frac{1}{C}\sum_{r=1}^{C} \exp\mathcal{L}(X|\theta_{\hat{r}})$. For VQ and AHS we approximate $\lambda = -d_{\hat{r}} + \frac{1}{C}\sum_{r=1}^{C} d_r|_{r\neq\hat{r}}$. We use the $C = 3$ closest cohorts[7].

3. RESULTS

We report identification error rate and verification equal error rate (EER) for the different models, features, compensation methods and test conditions. The relative errors should

[5] We obtained best results for $20 \leq p \leq 24$.

[6] This provides some robustness to changes in energy.

[7] We did not use the speakers that are in the cohort set of a claimant as imposters when reporting the equal error rate.

be interpreted with caution in that it is impossible to make 'all things equal' and any of the models and features can of course always be optimized more. The aim here is to give an indication of the behavior of the models under different conditions.

3.1. Experiments

We test the robustness of the analysis features and models for a population of 20 speakers[8] taken from the multi-environment database described in the introduction. The experiments are performed with a vocabulary of 10 words */processing abracadabra singularity nebula startrek supernova computer sungeeta generation tektronix/* from which we draw 7 unique words with 3 examples of each [21 utterances, $\sim$ 10 sec of speech] to train the models with, and use the remaining 3 words with 4 examples of each [12 utterances] for testing. This selection ensures that words used for testing were not used for training. To smooth error estimates we repeat this process for 3 different sets of test words and average the errors[9].

We perform two different tests 1) 4 tests per speaker with test words concatenated separately for each of the 4 examples [$\sim$ 1.3 sec of speech per test]) and 2) 1 test per speaker with test words concatenated for all 4 examples [$\sim$ 5 sec of speech per test]). The second test smoothes within-utterance variability. We use speech from 4 different environments: 1) office telephone (internal lines), 2) home telephone, 3) carbon-button microphone (internal lines) and 4) speaker telephone (internal lines). We train models individually for each environment and test individually on all environments.

Effect of environment. Table 1 lists percentage identification errors for training and testing within and across the four different environments. Within environments errors are relatively low, while across environments errors increase substantially. Since errors for the speaker telephone (4) is substantially worse than for the other environments and may bias conclusions we decided not to use it in subsequent experiments.

Effect of analysis method and model. Table 2 lists identification errors and Table 3 lists verification errors for the analysis methods and models. While errors are relatively low within environments, it is seen that when testing and training in different environments (the realistic telephone scenario) the errors rise substantially. PLP with RASTA processing appears to be the most robust of the different analysis methods.

GMM gives across analysis methods the lowest percentage error (which may be due to it being the best positioned to model complex interactions between the feature vectors).

[8] We used the remaining 16 speakers for cross-validation.

[9] For standard deviation estimates we assume that tests across word sets are independent, but that tests across environments are not.

Train, Test	1	2	3	4
1	4.6	32.9	27.1	35.8
2	45.0	5.0	37.9	35.0
3	27.9	37.9	10.4	32.5
4	57.1	50.4	49.2	4.6

Table 1: Average identification error (%) for the GMM classification method, and RASTA PLP analysis when training and testing individually in one of four environments. Tests used approximately 1 sec speech. (1=office telephone, 2=home telephone, 3=carbon-button microphone, 4=speaker telephone.) (Average standard deviation of the errors is 5.4, min=2.2, max=10.5)

It is however interesting to note the difference in effect of RASTA processing for AHS and GMM. RASTA is seen to affect GMM little, but results in a substantial improvement for AHS. (This is to be contrasted to the finding that GMMs easily outperform AHS on NTIMIT [2].) An explanation for this effect may be that the bandpass filtering of RASTA reduces spurious modalities in the data.

For verification we note that the ROC curves have large tails. Eg, for GMM and Rasta-PLP+CMS with 1 sec of test speech, training on environment (1) and testing on environment (2) the ROC has (false acceptance, false rejection) at (7.6%,45%), (23%,20%), (61%,7.5%).

	length	AHS	GMM	VQ
	(sec)	a) Same environment		
lpc	1	11.5(4.5)	7.8(4.2)	10.4(4.1)
lpc+r	1	10.0(5.4)	6.9(3.8)	10.6(5.1)
plp	1	7.5(4.0)	5.0(3.7)	8.6(4.1)
plp+r	1	4.4(3.2)	6.7(3.4)	9.4(4.0)
lpc	5	6.1(4.3)	4.4(3.4)	6.1(3.2)
lpc+r	5	4.4(4.7)	2.8(1.5)	5.0(2.9)
plp	5	5.0(3.2)	2.8(2.5)	4.4(3.0)
plp+r	5	2.8(2.9)	3.3(2.5)	4.4(2.3)
		b) Different environments		
lpc	1	49.7(4.9)	43.6(6.0)	48.7(4.7)
lpc+r	1	45.8(5.5)	42.3(5.3)	48.8(3.8)
plp	1	41.3(6.3)	38.1(6.5)	40.2(5.7)
plp+r	1	41.6(6.0)	34.7(6.6)	36.7(6.3)
lpc	5	43.0(5.9)	37.0(7.3)	40.6(5.7)
lpc+r	5	40.0(6.4)	33.1(6.3)	38.7(6.5)
plp	5	37.2(8.1)	31.7(7.4)	34.8(7.3)
plp+r	5	33.5(6.8)	29.8(7.5)	29.3(7.3)

Table 2: Average identification error (%) for the classification methods (AHS, GMM and VQ), features (plp and lpc) and RASTA compensation method (r=RASTA) when training and testing in a) the same environment and b) different environments. CMS is done in all tests. (Standard deviation of estimate is shown in parenthesis.)

	length	AHS	GMM	VQ
	(sec)	a) Same environment		
lpc	1	10.8(2.1)	9.9(2.8)	10.3(2.3)
lpc+r	1	9.9(2.9)	10.4(3.6)	10.2(3.2)
plp	1	9.3(1.7)	8.1(2.4)	7.3(1.8)
plp+r	1	8.0(2.0)	9.0(2.6)	9.1(2.5)
lpc	5	7.6(2.0)	7.9(2.7)	8.9(2.9)
lpc+r	5	7.8(3.3)	8.3(3.1)	8.0(3.5)
plp	5	7.9(2.0)	7.4(2.3)	5.6(1.7)
plp+r	5	6.1(1.9)	7.0(2.7)	7.1(3.1)
		b) Different environments		
lpc	1	25.0(2.0)	23.8(2.6)	23.2(2.1)
lpc+r	1	23.2(2.5)	22.9(2.0)	22.8(1.7)
plp	1	23.2(2.5)	22.4(2.8)	20.4(2.0)
plp+r	1	23.0(2.7)	18.6(2.4)	18.1(2.6)
lpc	5	23.4(2.1)	23.1(3.4)	20.3(2.7)
lpc+r	5	20.2(2.8)	21.6(2.5)	20.2(1.8)
plp	5	21.4(2.7)	21.1(3.5)	18.7(2.7)
plp+r	5	20.4(2.9)	17.3(2.9)	15.5(2.8)

Table 3: Average verification EER (%) for the classification methods (AHS, GMM and VQ), features (plp and lpc) and RASTA compensation method (r=RASTA) when training and testing in a) the same environment and b) different environments. CMS is performed in all tests. (Standard deviation of estimate is shown in parenthesis.)

Effect of compensation and normalization. Table 4 shows the effect of not doing cepstral mean subtraction. It is seen that when RASTA is not used errors improve when training and testing within the same environment, but worsen substantially when the environment is different. This indicates that although some speaker information is present in the cepstral mean it cannot be used reliably across environments.

	length	AHS	GMM	VQ
	(sec)	a) Same environment		
plp	1	3.3(2.4)	2.2(2.7)	3.9(2.2)
plp+r	1	4.6(3.1)	6.8(3.5)	9.4(4.0)
plp	5	1.1(1.3)	2.2(2.9)	2.2(1.5)
plp+r	5	2.8(2.1)	3.9(2.4)	4.4(2.3)
		b) Different environments		
plp	1	45.4(9.3)	47.6(8.7)	46.7(7.6)
plp+r	1	42.6(6.4)	35.0(6.6)	36.5(6.4)
plp	5	42.0(10.2)	46.1(9.8)	44.1(8.8)
plp+r	5	34.8(7.5)	29.6(7.7)	29.4(7.3)

Table 4: CMS not performed: Average identification error (%) for the classification methods (AHS, GMM and VQ), features (plp and lpc) and RASTA compensation method (r=RASTA) when training and testing in a) the same environment and b) different environments. (Standard deviation of estimate is shown in parenthesis.)

4. CONCLUSIONS

We conclude that PLP generally outperforms LPC. We also conclude that the statistical AHS measure can outperform the unsupervised clustering methods of VQ and GMM. This happens when the spectral trajectories are bandpass filtered and would motivate investigation of dynamic features (such as transitions at the syllabic rate.) In contrast, complex models such as GMMs give best results when this filtering is not performed. We therefore conclude that RASTA processing removes reduncancies from the features.

We affirm that speaker recognition is highly sensitive to acoustic mismatch in the telephone environment with errors high when training and testing across environments.

5. ACKNOWLEDGMENTS

This research supported in part by grants from the National Science Foundation, the Advanced Research Projects Agency, and the member companies of CSLU. The author would like to thank Todd Leen, Etienne Barnard and Hynek Hermansky for their suggestions.

6. REFERENCES

1. F. Bimbot and L. Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *Eurospeech*, pages 169–172, Berlin, 1993.
2. J.-L. Le Floch, C. Montacie', and M.-J. Caraty. Speaker recognition experiments on the NTIMIT database. In *Eurospeech*, pages 379–382, Sept 1995.
3. S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29:254–272, April 1981.
4. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, April 1990.
5. H. Hermansky. Rasta processing of speech. *IEEE Trans. Speech and Audio Processing*, 2(4), Oct 1994.
6. D. Mansour and Biing Hwang Juang. A family of distortion measures based upon projection operation for robust speech recognition. In *ICASSP*, pages 36–39, April 1988.
7. T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *ICASSP*, pages 157–160, 1992.
8. D. Ormoneit and V. Tresp. Improved gaussian mixture density estimates using Bayesian penalty terms and network averaging. In *Advances in Neural Information Processing Systems*, volume 8, page not yet available. The MIT Press, 1996.
9. D.A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Trans. Speech and Audio Processing*, 2(4):639–643, Oct 1994.
10. D.A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.
11. M. Schmidt, H. Gish, and A. Mielke. Covariance estimation methods for channel robust text-independent speaker identification. In *ICASSP*, pages 333–336, 1995.