TEXT INDEPENDENT SPEAKER IDENTIFICATION WITH CODED SPEECH

Ivan Kraljevski, Maria Paola Bissiri, Rüdiger Hoffmann

TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany
{ivan.kraljevski, maria_paola.bissiri, ruediger.hoffmann} @tu-dresden.de

Abstract. In this paper, a system for text independent speaker identification was evaluated under different coding conditions over limited speech data. The identification experiments were performed on the UASR (Unified Approach for Speech Synthesis and Recognition) system with two different configurations. Performance evaluations over different feature sets, number of Gaussians per model, as well as amount of training data were performed in order to select an appropriate configuration for speech coding effects demonstration. The initial results did not appear suitable for practical application and further improvements were achieved with model compensation by adaptation on different coders. In this case, significant improvements were observed even with limited amount of speech data with similar identification rates as for clean speech. In case of large speech data mismatch or low bit-rate coded speech, model training on coded speech was performed to further improve speaker identification.

1 Introduction

Beside its linguistic content, speech carries also unique information about the speaker, regarding anatomy (pitch and vocal tract resonances) and characteristic speaking manner (accent, rhythm, intonation style, pronunciation pattern, vocabulary etc.)[1]. These unique characteristics can be used in security systems as biometrical features for speaker recognition.

There are two possible applications of speaker recognition: speaker verification and speaker identification. Verification is used to confirm the claimed speaker identity based on his voice sample, while identification is the process of recognizing a speaker from a given database and deciding about positive or negative identification. If the person is required to speak a previously known utterance (prompt) during enrollment and recognition phase, the system is considered text-dependent, otherwise text-independent [2].

Speaker identification can be used in two operational modes: "open-set", where the speaker is part of the general population and "closed-set", where the speaker is identified as part of the existing speaker database. The main performance measure is the speaker identification rate, defined as the percentage of correct identifications averaged across all speakers in the database [3].

Speaker identification systems are usually based on Gaussian Mixture Models (GMM) and on the Universal Background Model (UBM) approach where the used features are the Mel-frequency cepstral coefficients (MFCC). For each speaker, a GMM model is created using the available training data. Speaker-dependent Gaussian components can represent general acoustic classes that reflect speaker-dependent vocal tract shapes and can model arbitrary densities [4].

Two factors significantly affect the performance of speaker identification systems: training and testing conditions mismatch and limited amount of available speech data. Differences in acoustic environment (noise, reverberation), technical recording conditions (transmission channel, microphone), as well as within-speaker variation (health, mood, aging) significantly affect identification rate [5]. Several methods have been developed to enhance the robustness of speaker identification systems for given acoustic environment and technical conditions [1].

They can be grouped into the following categories: feature compensation [6], model compensation [7] or re-scoring techniques.

Today, the Internet is the most popular and used medium for distributing information in various forms: texts, images, on-line and off-line audio and video content. With the global spread of Internet and of multimedia technology, an increasing number of audio signals – which are coded by some audio or speech coders, transmitted on the Internet or stored as multimedia files – is available. There is an increasing greater need of using speech and speaker recognition over multimedia content available on the Internet in various recognition tasks, as IVR services over VoIP, in mobile services, etc.

The authors in [5] investigate the effect of audio coding in speaker identification and verification in matched and mismatched testing and training conditions using popular audio coding algorithms on a system based on Gaussian mixture models. They reported slight decrease of performance in the case without sample rate change and significant loss when the sample rate was changed during audio coding. The degradation of recognition performance as reported for low bit-rate audio and speech coders is difficult to model and conventional noise canceling techniques as such as spectral subtraction, cepstral mean subtraction and RASTA, cannot be applied. For instance, the effect of GSM coding in the cepstral domain leads to a spreading and displacing of the means of the Gaussians [8]. The accuracy performance can be improved by either retraining of the speakers' models or using conventional adaptations methods in mostly cases. However, in order to perform any of these approaches, a large number of appropriately processed training/adaptation sentences to represent particular environmental or channel model influences should be provided [9].

In this paper, a system for text independent speaker identification was evaluated under different coding conditions over limited speech data. The identification experiments are performed on the UASR (Unified Approach for Speech Synthesis and Recognition) system with two different experimental setups. Performance evaluation over different feature sets, number of Gaussians per model, as well as the amount of training data were performed in order to choose a configuration for speech coding effects demonstration. Adaptation and matched training were also performed in order to assess whether further improvements are possible in the case of coded speech.

2 Influence of coding and model adaptation

2.1 Effects of coding

Speech signals are usually compressed by using lossy algorithms, which remove redundant information from the original signal, introducing distortion and reducing the size for transmission or storage. This has little effects on the perceptual quality, while it significantly deteriorates the accuracy of the speaker identification/verification system. Since the goal of speech coders is to maintain intelligibility of phonetic information, it is not clear how much speaker-dependent information is removed or degraded.

From the analysis presented in [10] it can be seen that the coding-decoding distortion can be modeled as a Gaussian density function. Based on empirical observations and clean and coded speech signals comparisons in this study, it can be suggested that the cepstral coefficient in frame *t* of the original signal $S^o_{t,n}$, can be represented as:

$$S^o_{t,n} = S^d_{t,n} + D_n \quad (1)$$

the S^d is the cepstral coefficient corresponding to the coded speech signal; D_n is the distortion caused by the coding-decoding process with probability density function:

$$f_{D_n}(D_n) = N(\mu_n^d, \Sigma_n^d) \tag{2}$$

where *N:*

$$N(\mu_n^d, \Sigma_n^d) \tag{3}$$

is a Gaussian distribution with a mean μ and variance Σ. Hence, the coding-decoding distortion is modeled as additive process in the cepstral domain, which implies the introduction of additive correction terms in the mean μ and variance Σ, independent from the phonetic class and from the output probability densities. Additive correction in the mean and variance parameters has been already successfully applied in the context of speaker adaptation [11]. Here compensation depends on the phonetic class and requires larger amount of adaptation data, which is the main issue in model adaptation. These observations justify the usage of conventional adaptation algorithms like MAP for speaker model adaptation in the case of coded speech.

2.2 MAP based coding compensation

Maximum a posteriori (MAP) [11] is a method for acoustic model adaptation which in the case of speaker recognition systems is used as an alternative for speaker model creation by adapting UBM on small speaker data. Although there is a risk of smearing the distinctive speaker's spectral characteristics, adaptation can be used as a model compensation technique also in case of mismatched channel conditions. In [12] an approach is presented where a channel independent model is transformed into a set of channel dependent models using the mapping parameters in the feature domain.

The MAP adaptation updates the Hidden Markov Model (HMM) parameters by joining known information (the old parameters) with the statistics derived from the adaptation data using data-dependent weighting coefficient. The data dependency is designed to weight the statistics with higher population towards new parameters and the statistics with lower population towards the original parameters. The new mean and covariance (4) for the distribution *j* presents weighted sum of the old and the new statistics:

$$\hat{\mu}_j = \frac{n_j}{n_j + \rho_\mu}\tilde{\mu}_j + \frac{\rho_\mu}{n_j + \rho_\mu}\mu_j, \quad \hat{\Sigma}_j = \frac{n_j}{n_j + \rho_\Sigma}\tilde{\Sigma} + \frac{\rho_\Sigma}{n_j + \rho_\Sigma}\Sigma_j \tag{4}$$

The data-dependency of the weighting coefficients is realized by the count of the adaptation data n_j and the relevance factors ρ_μ and ρ_Σ. Their values mark the points where the new parameter has the same weight as the old one. Higher values of ρ give more weight to the prior information, the old parameters. The main problem with the MAP adaptation is that it is an unconstrained method and updates therefore only those parameters for which observations exist. It requires a relatively large amount of adaptation data in order to be effective for sparsely occupied Gaussian distributions.

3 Experimental setup

3.1 The identification system

The used identification system is based on a maximum-likelihood classifier. For a reference group of *N* speakers $S = \{1, 2, \ldots, N\}$ represented by models $\lambda_1, \lambda_2, \ldots, \lambda_N$, the objective is to find the model with the maximum posterior probability for the given utterance feature vector sequence, $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$ using Bayes decision rule:

$$\hat{S} = arg \max_{1 \le k \le N} P_r(\lambda_k | X) = arg \max_{1 \le k \le N} \frac{P(X|\lambda_k) P_r(\lambda_k)}{p(X)} \quad (5)$$

where $P_r(\lambda_k)$ is the prior probability for speaker model λ_k, $p(X)$ is the prior probability for the utterance feature vector sequence X and $P(X|\lambda_k)$ is the likelihood of feature vector sequence X to correspond to the model λ_k. Assuming equal speaker prior probability, the terms $P_r(\lambda_k)=1/N$ and $p(X)$ are constant for all speakers and can be omitted. Using logarithms and the assumed independence between the observations, the decision rule becomes:

$$\hat{S} = arg \max_{1 \le k \le N} \sum_{t=1}^{T} \log p(\vec{x}_t | \lambda_k) \quad (6)$$

where $p(\vec{x}_t | \lambda_k)$ is the output probability for $\vec{x}_t$ to match speaker model λ_k.

The used identification system was implemented by modified UASR (Unified Approach for Speech Synthesis and Recognition) framework. Basically, the system uses arc-emission HMMs with single Gaussian density per arc and an arbitrary topology. The structure is built by iterative training process by means of state splitting from an initial HMM models [13].

For the purpose of speaker identification it was used in two different experimental configurations. In the first one, individual speakers were modeled by 3-state HMMs, in the second one with single state HMMs (GMM). The speaker models are built on clean and coded speech by state splitting, effectively increasing the Gaussians number exponentially on power two of the split iteration (after split 1 there are two, after split 2 - four and so on).

During the feature extraction process, the clean and coded-decoded speech signals, with sampling frequency of 16 kHz and 16 bits resolution, were divided in 32 ms wide frames with a shift of 10 ms and processed with a Hamming window. The band from 300 to 8000 Hz was covered with 31 Mel DFT filters and at the output of each channel the log of the energy was computed. The obtained feature vectors and theirs delta values were standardized to a mean of zero and a standard deviation of one.

In several experiments, feature extraction process was combined with Principal Component Analysis (PCA). The effectiveness of PCA in pattern recognition lies in its ability to de-correlate feature parameters and relegate most of the random structures to trailing components while extracting systematic patterns to leading ones. It is assumed that PCA could aid the robustness in the case of speaker identification over coded speech.

3.2 Databases and speaker modeling

The used speech database consisted of studio recorded sequences with Microtech Gefell M930 microphone, processed and formatted with 16 kHz and 16 bit PCM quality. Twenty speakers (11 male, 9 female) participated in single, one hour sessions, producing on average 20 min speech material per person. The larger training set consisted of approximately 70 sentences (~14 min) and the test set of 30 sentences (~6 min) per speaker. Limited training (2,5 min) and test sets (25 sec) per speaker were also prepared.

The speaker modeling was performed under two different experimental setups. In the first, for each speaker, HMM models (3 states) were used along with silence (3 states) and Universal Background Model (UBM) with 6 states. In this case, UBM was used in "open-set" speaker identification for the detection of unknown speakers (imposter model), however in closed-set speaker identification systems there is no need for a UBM since the individual speaker GMMs are sufficient to carry out the identification process. In the second setup, single state models (GMMs) were used for speaker modeling without the UBM and silence model.

The models were trained with different number of Gaussian density distributions: 1, 2, 4, 8, 16, 32 and 64 per state. Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPC) and Melfilter (MEL) features in combination with Principal Component Analysis (PCA).

4 Identification experiments on UASR

Performance evaluation over different types of features, number of Gaussians per state, as well as the amount of training data were performed in order to choose a suitable configuration to demonstrate the effects of speech coding. The first set of experiments were carried out on a HMM based identification system with 3 states per speaker model (and silence) and 6 states for UBM, using MFCCs, covariance matrices, closed speaker set with large (noted as EXP 1-1) and small (EXP 1-2) amount of speaker data.

The best identification results in both cases (large and small amount of data) were achieved with 16 Gaussians per state, indicating that there is a lower limit for the number of mixture components to model individual speakers [4].

For the larger training and test data the highest achieved accuracy was 92.1% and for the limited speech amount 75.0% (s. Table 1), indicating that this setup is not appropriate in cases when limited speech data is available. There were also increased confusions between UBM and the speaker models due to the limited amount of training data for reliable UBM training.

In the second set of experiments GMM speaker models were used, without silence and UBM models, with MFCC features and only variances, on closed speaker set for large (EXP 2-1) and limited (EXP 2-2) speech data. The silence model was omitted and silence periods were not labeled, assuming that with the increase of mixture numbers silence will be accordingly represented within the speaker models. The identification rates improved (99.8% - large and 100% - limited training set) as a result of the reduced number of models and the exclusion of UBM.

Table 1. Speaker identification rate (%) over number of state splits

Experiment/Model	0	1	2	3	4	5	6
EXP 1-1	90.1	91.3	91.9	91.9	92.1	91.9	91.1
EXP 1-2	71.4	73.2	73.2	75.0	75.0	71.4	71.4
EXP 2-1	96.8	95.3	88.7	99.5	99.8	99.8	100
EXP 2-2	97.6	97.6	81.0	92.9	100	100	100

It was also observed that increasing the number of Gaussian mixtures does not necessary result in better identification (it was higher for models with 1 and 16 mixtures than for the model with 4). The reason is, although the model with one mixture is roughly approximated in the feature space, that the distance between the mean values is sufficiently large for good speaker identification, which was not the case with models with 4 mixtures. This configuration was used for identification experiments across different feature extraction methods (MFCC+PCA, MEL+PCA, LPC+PCA) (Table 2) and different speech coders (Table 3).

Table 2. Speaker identification rate (%) over different features and number of state splits

Experiment/Model	0	1	2	3	4	5	6
MFCC+PCA	92.9	90.5	78.6	88.1	95.2	97.6	97.6
MEL+PCA	90.5	90.5	54.8	88.1	100	97.6	97.6
LPC+PCA	90.5	88.1	83.3	83.3	92.9	97.6	95.2

4.1 Coder types used in M-C training and adaptation

For the purpose of speaker identification over coded speech, nine different coders were chosen by their bit-rate coverage of the coding standards, their common use in multimedia and mobile applications as well as their availability and licensing. The speech coders: ADPCM, G.722, G.726, GSM, LPC-10 and the general audio coders: MP2, Vorbis, Real 14.4, WMA1 and WMA2. FFMPEG (open source) [14] and SOX (open source) [15] were used as a general coding-decoding engines.

4.2 Mismatched data experiments

Identification rate evaluation was performed over different coder types and bit-rates using the GMM based system (as in EXP 2-2), with different feature set combinations (Table 3). In all experiments the used speakers' GMM models were created after the fourth split, which means 16 Gaussians per model.

Table 3. Speaker identification rate (%) over different coders and feature sets

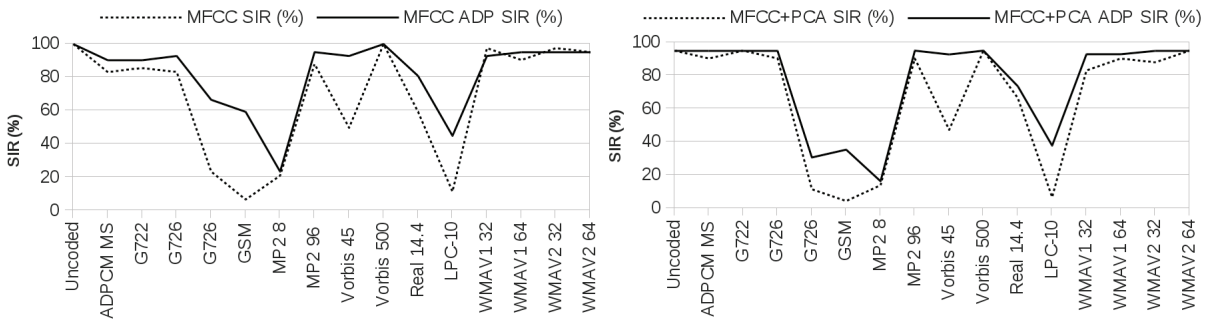
Type	CODER	Bit Rate kbps	MFCC (large data)	MFCC (limited)	MFCC+PCA (limited)	MEL+PCA (limited)	LPC+PCA (limited)
General	Uncoded	256	**100.00**	**100.00**	**95.20**	**100.00**	**92.90**
General	ADPCM	32	66.00	83.30	90.50	64.30	19.00
Voice	G722	64	83.30	85.70	95.20	71.40	21.40
Voice	G726	64	78.60	83.30	90.50	64.30	26.60
Voice	G726	32	35.70	23.80	11.90	19.00	14.30
Voice	GSM	13.2	14.30	7.10	4.80	14.30	19.00
General	MP2	8	23.80	21.40	14.30	16.70	4.80
General	MP2	96	90.50	88.10	90.50	78.60	9.50
General	Vorbis	45	50.00	50.00	47.60	73.80	9.50
General	Vorbis	500	100.00	100.00	95.20	100.00	92.90
Voice	Real14.4	14.4	54.80	59.50	66.70	28.60	38.10
Voice	LPC-10	2.4	19.00	11.90	7.10	21.40	16.70
General	WMAV1	32	97.60	97.60	83.30	92.90	23.80
General	WMAV1	64	90.50	90.50	90.50	90.50	14.30
General	WMAV2	32	100.00	97.60	88.10	95.20	28.60
General	WMAV2	64	97.60	95.20	95.20	100.00	19.00
Average			66.78	66.33	64.76	62.07	23.83

There is a significant decrease in the identification rate for the low bit-rate speech coders (GSM, LPC-10, MP2 - 8 kbps and Real 14.4). GMM models with 16 mixtures and MFCC features, alone and combined with PCA, produced the best results (average 66.33% and 64.76%) in coding conditions on a limited amount of data. However, these results are not suitable for practical applications and further improvement should be achieved using feature or model compensation techniques.

4.3 Adapted model experiments

Although MAP adaptation as a GMM model compensation method might introduce loss of speaker discriminative information, the trained models were adapted on a specific coder. The difference in speaker identification rate was observed and compared with the baseline model for the feature sets MFCC (Figure 1a) and MFCC combined with PCA (Figure 1b).

In both cases, significant improvements and satisfying identification rates were observed (particularly for Vorbis with 45 kbps), except in the cases in which the speech data mismatch was too large (GSM, MP2 8 kbps, LPC-10, etc.). There is a larger relative improvement in the case of MFCC alone than in the case of MFCC+PCA after adaptation for the low-bit rate coders except for MP 2 with 8 kbps.



a) b)

Figure 1. Speaker identification rate for clean and adapted model, a) (MFCC) and b) (MFCC+PCA)

4.4 Matched data experiments

For the low speech quality coders, matched training was performed in order to see whether it is possible to reach appropriate identification performance comparable between clean and adapted speaker models. The results presented for MFCC (Table 4) and MFCC+PCA (Table 5) exhibit similar behavior as in the case with clean speech. There is a lower limit of number of Gaussians per model that are sufficient for speaker modeling.

Table 4. Speaker identification rate (%) in case of matched training for MFCC features

GMM MFCC /split	0	1	2	3	4	5	6
Clean	97.6	97.6	81.0	92.9	100	100	100
GSM	88.1	88.1	81.0	90.5	97.6	97.6	100
MP2 8Kbps	85.7	88.1	61.9	66.7	90.5	88.1	95.2
LPC-10	90.5	88.1	83.3	88.1	92.9	95.2	95.2
Real 144	95.2	92.9	83.3	78.6	85.7	90.5	95.2

Table 5. Speaker identification rate (%) in case of matched training for MFCC+PCA features

GMM MFCC+PCA /split	0	1	2	3	4	5	6
Clean	92.9	90.5	78.6	88.1	95.2	97.6	97.6
GSM	88.1	88.1	81.0	92.9	95.2	97.6	97.6
MP2 8Kbps	85.7	85.7	47.6	85.7	95.2	95.2	95.2
LPC-10	92.9	92.9	88.1	92.9	95.2	95.2	95.2
Real 144	85.7	85.7	76.2	83.3	90.5	92.9	90.5

For the MFCC+PCA feature set, after fourth split, for the most of the used coders, no further improvements could be observed while increasing the number for Gaussians. More mixtures per model are needed to maximize the speaker identification rate in the case of MFCC features.

5 Conclusions

In this paper we investigated the performance of text-independent speaker identification over speaker models trained on clean speech, models adapted and models trained on coded speech. Even with limited amount of speech data, a similar identification accuracy as for clean speech could be achieved using adaptation for high quality coders. In the case of low bit-rate coded speech, significant improvements in speaker identification rate could be achieved by means of training in matched conditions, for example for GSM coded speech with: a) MFCCs, on clean models: 7.1%, on adapted models: 59.5%, and on matched trained models: 97.6%; and

b) MFCCs+PCA on clean models: 4.8%, on adapted models: 35.7%, and on matched trained models: 92.9%. Future work will be focused on using feature-based combined with model-based compensation, as well as including better UBM modeling and testing the system on standardized speakers databases.

References

[1] Kinnunen, Tomi, and Haizhou Li. "An overview of text-independent speaker recognition: From features to supervectors", Speech Communication 52, no. 1 (2010): 12-40.

[2] Togneri, Roberto, and Daniel Pullella, "An overview of speaker identification: Accuracy and robustness issues". Circuits and Systems Magazine, IEEE 11, no. 2 (2011): 23-61.

[3] A. R. Stauffer, A. D. Lawson, "Speaker Recognition on Lossy Compressed Speech using the Speex Codec", Proceedings of the Interspeech 2009 Brighton United Kingdom, 6-10 September-2009.

[4] Reynolds, Douglas A. "Speaker identification and verification using Gaussian mixture speaker models". Speech communication 17, no. 1 (1995): 91-108.

[5] Jiang, Tao, Boyang Gao, and Jiqing Han. "Speaker identification and verification from audio coded speech in matched and mismatched conditions". Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on. IEEE, 2009.

[6] J. Pelecanos, S. Sridharan, "Feature Warping for Robust Speaker Verification", ICSA Archive, in A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, 18-22.6.2001, pp. 213-218.

[7] R. Teunen, B. Shahshahani, L. Heck, "A model-based transformational approach to robust speaker recognition", Proc. of ICSLP, pp.495-498, 2000.

[8] J. M. Huerta, "Speech recognition in mobile environments", Ph.D dissertation, Dept. Elect. Comput.Eng., Carnegie Mellon Univ., Pittsburgh, PA, Apr. 2000.

[9] Yamada, Miichi, et al. "Unsupervised acoustic model adaptation algorithm using MLLR in a noisy environment". Electronics and Communications in Japan (Part III: Fundamental Electronic Science) 89.3 (2006): 48-58.

[10] Yoma, Néstor Becerra, Carlos Molina, Jorge Silva, and Carlos Busso. "Modeling, estimating, and compensating low-bit rate coding distortion in speech recognition". Audio, Speech, and Language Processing, IEEE Trans. on 14, no. 1 (2006): 246-255.

[11] J. L. Gauvain, C-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", in IEEE Transactions on Speech and Audio Processing, Vol.2, No.2, April 1994, 291-298.

[12] D. A. Reynolds, "Channel robust speaker verification via feature mapping", in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP), 2003, vol. 2, pp. 53–56.

[13] R. Hoffmann, M. Eichner, and M. Wolff: "Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system". In: A. Esposito et al. (eds.), Verbal and Nonverbal Communication Behaviors. Berlin etc.: Springer 2007, Lecture Notes in Artificial Intelligence vol. 4775, pp. 200-218.

[14] ffmpeg.org, 07.01.2013

[15] sox.sourceforge.net, 05.03.2012