PHONE-BASED SPEAKER VERIFICATION WITH VARIOUS ADAPTATION CONFIGURATIONS

Yossi Bar-Yosef and Yuval Bistritz

Department of Electrical Engineering
Tel-Aviv University, Tel-Aviv 69978, Israel
yossibaryosef@gmail.com, bistritz@eng.tau.ac.il

ABSTRACT

This paper introduces an examination of several adaptation schemes for phone-based speaker verification. It is argued that different phonemes convey different amount of valuable information about speaker classification. Thus, the presented study tries to find more optimal settings to exploit this information. Experiments with short duration of training and testing of clean and telephone text-independent speech highlighted the superiority of one configuration over the rest. This configuration implements a two-stage adaptation of phone models, outperforming the standard phone-independent GMM-based system. The paper also considers adaptation of a subset of the whole phonetic set, and its comparative improvement of performance. Experiments showed that this partial adaptation approach, beyond reducing storage requirement, significantly improves the overall performance.

Index Terms— Speaker verification, Gaussian Mixture Models, Speaker phoneme model adaptation.

1. INTRODUCTION

The paper considers text independent phone-based speaker verification. Speaker verification (SV) is a branch of speaker recognition where the machine has to accept or reject the claimed identity of a speaker given a sample of his speech. The most common approach to text-independent speaker recognition today is using Gaussian Mixture Models (GMMs) [1]. A GMM-based SV system was found in recent evaluations to be the best performer [2]. Even though in these systems speakers are modelled each by a single GMM, it has been observed that the Gaussian components tend to represent the underlying phonetic sounds of a speaker's voice. Speaker verification methods based on Hidden Markov Models (HMMs) proposed in [3] combine phonetic modelling of speech with temporal information. Auckenthaler et al. [4] compared a phone-independent approach based on GMMs and phone-based approach using Hidden Markov Models (HMMs). In all their experiments, the phone-independent GMM system has consistently outperformed the phone-based HMM system. However, they observed that the addition of phonetic weighting, borrowed from the segmentation that was carried out by the HMM system, improved the performance of their phone-independent GMM system. Newman et al. [5] which perform speaker verification through large vocabulary continuous speech recognition (LVCSR), report that this is a competitive alternative to available GMM SV systems. Phone-based speaker verification typically involves two stages. First, the speech is segmented to phone classes. At the second stage, given phone-dependent models for each speaker, the verification task is carried out by scoring each frame with its corresponding phone model. The verification procedure may be applied in several ways. Olsen used in [6] phone-dependent radial basis function neural networks, and considered different feature presentation for each phone for improved performance. In [7], a two-staged SV system was considered where speech was segmented to 8 different speech classes and, correspondingly, each speaker was assigned 8 acoustic GMMs. The segmental GMM system could not outperform the equivalent global GMM system. Jin et al. [8] developed a multilingual speaker identification system where the phone strings were derived from eight different languages phone recognizers. The multilingual approach was found to be powerful for speaker identification, especially under non-matching conditions. Yet, it was noted that good performance was achieved only for large amount of training data. In a previous closely related study, D. Gutman and Y. Bistritz [9] generated phone-dependent GMMs for the target speaker by applying adaptation of the speaker's general GMM, and phone-dependent background GMMs by adapting the general background speakers' GMM. This configuration performed better than a standard GMM-based system when small size models were used. This paper proposes and examines several phone-dependent configurations based on various adaptation paths. Verification experiments were held on both clean and telephone speech using small amount of training and testing data. The systematic study reveals one of the several new configurations, a certain doubly adapted scheme, as best performer. The paper also examines adaptation of only subsets of the whole set of phones. A *Phone Knockout Rejec-*

Authorized licensed use limited to: TEL AVIV UNIVERSITY. Downloaded on January 20, 2009 at 11:04 from IEEE Xplore. Restrictions apply.

tion procedure is used to select the most effective subsets of phones. The experiments showed that this partial adaptation approach, beyond reducing storage requirement per target speaker, significantly improves the overall performance.

2. PHONE-BASED SPEAKER VERIFICATION

2.1. MAP Adaptation of GMMs

For a D-dimensional feature vector, x, a probability density of a GMM is defined as a weighted sum of M gaussian densities:

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{1}$$

where $p_i(x)$ is a unimodal Gaussian density, parameterized by a mean vector μ_i, , a covariance matrix Σ_i, and mixture weights w_i that add to unity. The model, λ, is collectively denoted as $\lambda = \{w_i, \mu_i, \Sigma_i\}$, where $\quad i = 1, \ldots, M$. Maximum a Posteriori (MAP) adaptation approach is used to update the model parameters to a new data. The MAP adaptation introduced by Gauvain and Lee in [10], relays on the assumption that the old model is well trained and that a supervised adjustment to the new data is required. The MAP adaptation is obtained in two steps. In the first step the new sufficient statistics are estimated. In the second step the new statistic estimates are combined with the old parameters using a data-dependent mixing. Given a GMM with diagonal covariances, λ, and training vectors $X = \{x_1, \ldots, x_T\}$, the probabilistic alignment of the training vectors into the mixture component i is computed as

$$Pr(i \mid x_t, \lambda) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} , \tag{2}$$

where $p_i(x_t)$ is the Gaussian density of the mixture component i given vector x_t. Then $Pr(i \mid x_t, \lambda)$ is used to compute the weight, mean, and variance:

$$\begin{aligned} n_i &= \sum_{t=1}^{T} Pr(i \mid x_t, \lambda) \\ E_i(x) &= \frac{1}{n_i} \sum_{t=1}^{T} Pr(i \mid x_t, \lambda) x_t \\ E_i(xx') &= \frac{1}{n_i} \sum_{t=1}^{T} Pr(i \mid x_t, \lambda) x_t x_t' , \end{aligned} \tag{3}$$

Next, the new statistics for mixture i is used to create the adapted parameters for the mixture i by combining the original parameters with the estimated parameters as follows:

$$\begin{aligned} \hat{w}_i &= [\alpha_i^{(w)} n_i / T + (1 - \alpha_i^{(w)}) w_i] \gamma \\ \hat{\mu}_i &= \alpha_i^{(m)} E_i(x) + (1 - \alpha_i^{(m)}) \mu_i \\ \hat{\sigma}_i^2 &= \alpha_i^{(v)} diag(E_i(xx')) + (1 - \alpha_i^{(v)})(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2 , \end{aligned} \tag{4}$$

The adaptation coefficients $\{\alpha_i^{(w)}, \alpha_i^{(m)}, \alpha_i^{(v)}\}$ control the balance between the old and new estimates, and γ is a scaling factor ensuring that all weights sum to unity. For each parameter and each mixture, $\{\alpha_i^{(\rho)}\},\ \rho \in \{w, m, v\}$, is the data-dependent adaptation coefficient, which is defined as:

$$\alpha_i^{(\rho)} = \frac{n_i}{n_i + r^{(\rho)}} , \tag{5}$$

where $r^{(\rho)}$ is a fixed relevance factor for parameter ρ. We used a single adaptation coefficient, $\alpha_i = \alpha_i^{(w)} = \alpha_i^{(m)} = \alpha_i^{(v)} = \frac{n_i}{n_i + r}$.

2.2. GMM-Based Speaker Verification

Reynolds et al. presented in [1] a GMM-based system in which a single Universal Background Model (UBM) is used to represent the alternative hypothesis. The UBM is a large GMM trained on a large pool of speakers' speech representing the speaker-independent feature distribution. The target speaker model is derived by adapting the parameters of the UBM using the speaker's training data. At testing, for a sequence of T feature vectors, $X = \{x_1, \ldots, x_T\}$, where λ_{hyp} is the model that characterizes the hypothesized speaker and $\lambda_{\overline{hyp}}$ characterizes the alternative hypothesis, the log likelihood ratio (LRT) is calculated as

$$\Lambda(X) = \log p(X \mid \lambda_{hyp}) - \log p(X \mid \lambda_{\overline{hyp}}) \tag{6}$$

where

$$\log p(X \mid \lambda) = \sum_{t=1}^{T} \log p(x_t \mid \lambda) \tag{7}$$

Then $\Lambda(X)$ is normalized by dividing it by T, and compared to a decision threshold which is set to adjust the tradeoff between rejecting true claimant utterances (False Reject errors) and accepting impostor utterances (False Accept errors).

2.3. Phone-Based Modelling and Testing

The first stage of our phone-based speaker recognition system, both at training and at testing, involves phonetic segmentation of the speech. At training, the entire collection of features extracted from the speech frames, X, are classified into K clusters of phones, $X_1, \ldots, X_K$ and each cluster is used for training a phone model. Segmentation is carried out identically at training and at testing. The speaker's phone-GMMs are generated using MAP adaptation which is most suitable to use when having small amount of training data. Each phone cluster is used for adapting a well-trained GMM to a phone-dependent GMM. We defined several different configurations of phone-based adaptation of GMMs. The adaptation is always carried out in the technique described in section 2.1. However, unlike the standard phone-independent system, where adaptation is used only for adapting a universal

Authorized licensed use limited to: TEL AVIV UNIVERSITY. Downloaded on January 20, 2009 at 11:04 from IEEE Xplore. Restrictions apply.

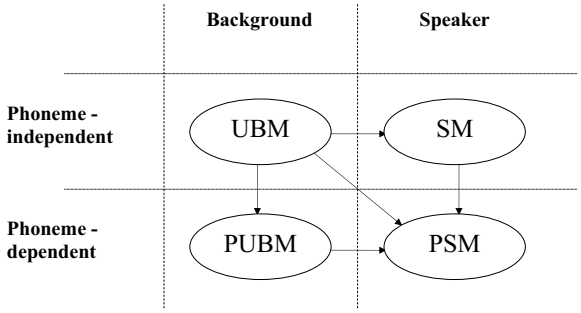

Fig. 1. Models categorization and adaptation schemes for speaker verification.

background model to the data of each speaker, in the following configurations the procedure is used in more subtle ways and at times more than once.

To begin with, we partition the models into four categories differentiated by their relation: *Background* or *Speaker*, and by dependency: *phone-independent* or *phone-dependent*. The 4 model categories are illustrated in Figure 1: 1) Universal Background Model (UBM) is a large GMM trained once and used for all target speakers in the verification task. 2) Speaker Model (SM) is the speaker-dependent GMM generated by using all the features from the training data of a specific speaker. 3) Phone-dependent Universal Background Model (PUBM) represents the speaker-independent distribution of features related to a specific phone. 4) Phone-dependent Speaker Model (PSM) represents the distribution of features related to a specific phone of a specific speaker.

Next, we focus on five configurations of phone-adapted modelling which are defined in Table 1. In the remaining of the paper they are being referred to as *Cfg1* to *Cfg5*, for brevity. The adaptation path is denoted by the "operator" $\rightarrow$.

Cfg. Name	Backgroung Modelling	PSM Modelling
Cfg1	$UBM \rightarrow PUBM$	$SM \rightarrow PSM$
Cfg2	UBM	$UBM \rightarrow PSM$
Cfg3	Direct $PUBM$	$PUBM \rightarrow PSM$
Cfg4	$UBM \rightarrow PUBM$	$PUBM \rightarrow PSM$
Cfg5	UBM	$UBM \rightarrow SM \rightarrow PSM$

Table 1. Five system configurations.

The verification procedure consists of two stages. First, the test utterance frames are segmented into phone segments by the speech recognition module. In the second stage the log likelihood is calculated for each frame using its correspondent phone adapted model of the hypothesized speaker. All frame scores are summed together and normalized by the total frame number. In the configurations that use PUBM ($Cfg1$, $Cfg3$ and $Cfg4$), the background score is calculated in the same manner. When using UBM as a background model ($Cfg2$ and $Cfg5$), the frame classification is not relevant.

3. EXPERIMENTS AND RESULTS

For phonetic segmentation we used the speech Recognition Experimental System (RES)[11] with 39 mono-phones.

The feature vector for speaker verification consisted of 12 mel-cepstrum coefficients, 12 delta mel-cepstrum coefficients, and a delta log-energy coefficient. For the telephone speech, cepstral analysis was performed over mel-filters in the pass-band 300-3400 Hz. For each of the configurations, verification was performed using several model orders, on a logarithmic scale: 4 to 128, except for *Cfg3* which was tested with model sizes of 1 to 16. A single adaptation rate parameter for the weights, means, and variance, with a relevance factor of $r = 12$ is used. Experiments were conducted on the phonetically-rich databases TIMIT (clean speech) [12] and NTIMIT (includes utterances recorded in TIMIT, passed through actual telephone lines) [13]. Using those databases allows us to consider the same speech samples differentiated only by transmission environments. 88 male speakers are used to train the Universal Background Model (UBM) and 350 male speakers were trained as target speakers. We present results in which training was done on approximately 20 seconds of speech. Testing trial was performed with approximately 3 seconds of speech, 2 true trials and 3 imposter trials for each target speaker.

In Table 2 we present the performance of the five phone-based system configurations by their Equal Error Rate (EER) measure, each configuration with the best performing model order. *UBM-SM* configuration represents the standard phone-independent system. Configuration *Cfg1* could not perform well in high order models, demonstrating the problem of adapting a speaker's model which is not well trained due to insufficient data. Generating phone-dependent speaker models by adapting the large UBM, as in *Cfg2*, dose not result in good performance. In contrast to *Cfg2*, the configurations *Cfg4* and *Cfg5* use a two-stage adaptation. In *Cfg4*, phone-dependent speaker models are generated from models that already **discriminate phones**. In *Cfg5*, the phone-dependent speaker models are generated from models that already **discriminate speakers**. They both perform better than *Cfg2*. The best performing configuration is *Cfg5*, which consistently outperforms the phone-independent system. An improvement of 31% was measured over the baseline configuration for clean speech, and 14% improvement for telephone speech.

The following observations characterize the results reported in this section: 1) Configurations, in which the phone-dependent speaker models were generated through an adap-

Authorized licensed use limited to: TEL AVIV UNIVERSITY. Downloaded on January 20, 2009 at 11:04 from IEEE Xplore. Restrictions apply.

Cfg.	Clean Speech		Telephone Speech	
	Order	EER (%)	Order	EER (%)
UBM-SM	128	1.6	128	8.0
Cfg1	16	1.9	8	10.6
Cfg2	8	2.6	32	10.3
Cfg3	8	1.7	16	7.6
Cfg4	4	1.7	64	7.9
Cfg5	128	1.1	128	6.9

Table 2. EERs (in %) in clean and telephone speech.

tation path starting from the background models, provide superior performance over a decoupled configuration where speaker model is trained independently, applying that the coupling between the background model and the target mode is most important. 2) Phone-based adaptation of a speake model should start with a well-trained GMM. Otherwise, this will lead to a worse phone-dependent modelling, resulting in a severe degradation in performance. 3) A significant improve ment in performance was measured using a new approach o double-stage adaptation path. This approach superimposes the benefits of the two principles observed above: (i) It generates a phone-independent speaker model by adapting the UBM to the speaker's speech, thus maintaining the "coupling" between the background model and the target model; (ii) It starts with an already well-trained phone-independent speaker GMM and adapts it to the phone data of that target speaker. The second adaptation stage provides a robust fine-tuning of the GMM's parameters to represent more accurately each speaker's phone.

4. PHONE SELECTION FOR SPEAKER VERIFICATION

Different phonetic classes carry different amount of useful information to speaker discrimination and have different impact on the speaker verification task. Using only a subset of the most discriminative phones while reducing or even omitting the contribution of the remaining phones has been shown to improve performance. For example, in [4] it was found that considering only a subset of 10 to 15 phones provides better results than using the whole set of phones. Performance of subsets of phones were previously considered also in [9].

We defined a selection procedure referred to as the *Knockout Rejection procedure*, in which we start with the full set of phone classes and then apply an iterative procedure for rejecting phones. The phone without which we obtain the best subset is discarded in each iteration. The *Knockout rejection procedure* requires $\frac{N(N+1)}{2} - 1$ evaluations. In our case ($N = 38$, excluding silence), 740 evaluations were conducted for each system configuration. This procedure is far more optimal than a simple selection of the N-best phones, and produced substantially better results. The effort invested

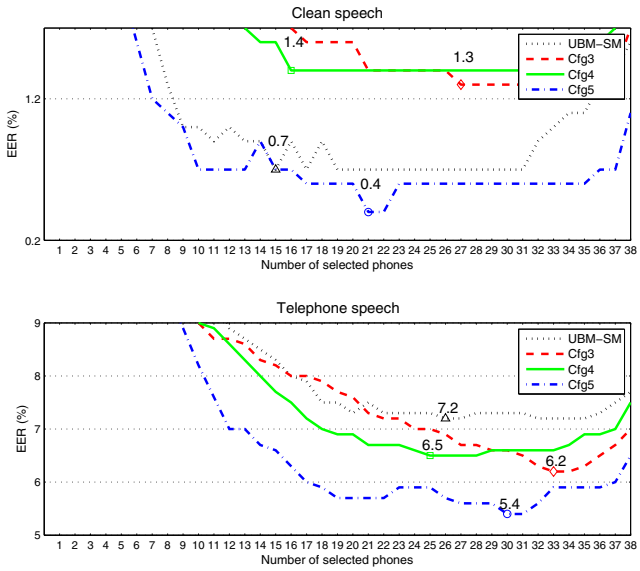

Fig. 2. EER (in %) using subsets of phones.

in applying this procedure is justified since the process is to be carried out once, before launching an operational system.

Phone selection results are presented in Figure 2 , showing the EER (in %) achieved by a subset of n selected phones. In clean speech, the best subset selected was consistently smaller in size than the best subset selected in the telephone environment. Furthermore, the percentage of improvement achieved by the selection procedure was considerably higher in clean speech. This result implies that a significant amount of speaker-discriminative information found in the higher frequency band (above 3400Hz). The lose of this information in telephone environment causes the effect of having more phones with little high band speaker-discriminative information contributing to the overall performance. Phone classes that were damaging the overall performance in the case of clean speech turned out to be contributing to performance improvement in the case of telephone speech. Hence, a larger subset of phones was selected to obtain best performance, and still, the improvement percentage in telephone speech was relatively smaller. Results showed that the most valuable phones in all system configurations are /n/ and /s/, in both speech environments. On the other hand, we observed the changes in phone importance when considering telephone speech instead of clean speech. For example, the phones /z/, /th/, and in some cases other fricatives have lost of their importance in favor of /ih/ and /iy/. This outcome is not surprising since fricatives have considerable information in the higher frequency band, which is lost in telephone environment.

The configuration that turned out to be using the larger subsets of phones was *Cfg3*. In this case, almost all the phones

Authorized licensed use limited to: TEL AVIV UNIVERSITY. Downloaded on January 20, 2009 at 11:04 from IEEE Xplore. Restrictions apply.

contribute to the total score, implying that this modelling technique is more robust in modelling each phone class parameters. But still, its mean performance is lower in the case of sparse training data. The configuration that achieved the top improvement by applying the phone selection procedure was *Cfg5*, in both cases of clean and telephone speech (64 % and 17% respectively). In *Cfg5* larger subsets of phones were used for the verification task in comparison to the standard *UBM-SM* configuration. This observation implies that the modelling technique of *Cfg5* improves the modelling of the speaker's phonetic sounds in a way that more phonetic clusters can contribute to the overall score computed for the verification task. The overall improvement in the EER of *Cfg5* with phone selection over the baseline configuration was 75% in clean speech and 30% in telephone speech.

5. CONCLUSIONS

The paper considered phone-based speaker verification with various adaptation configurations. Experiments on clean and telephone speech with short duration training and testing utterances were held. The best performing setting, was a configuration consisted of doubly-adapted phone models, that consistently outperformed the phone-independent GMM-based system. In this configuration, first a general target speaker model was obtained by adapting the parameters of the universal background model to the speaker's speech. Next, finely-tuned phone models were obtained by adapting the speaker's general GMM to his/her available clusters of phone data. A subsequent part of the paper studied the use of smaller sets of phones to improve performance. The paper has demonstrated clearly that a well designed phone-dependent GMM speaker verification system outperforms a comparable regular phone-independent systems. At the same time, phone-based speaker verification system is more expensive than a phone-independent standard system in terms of storage and computation. Storage restrictions may be partially alleviated by using only a subset of phone models. Also, the phone recognition layer does not pose difficulty when speaker verification is incorporated in an automatic speech recognition system or in applications that involve prompted text.

6. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, vol. 10, pp. 19-41, Jan. 2000.

[2] A. Martin and M. Przybocki, "The NIST Speaker Recognition Evaluations: 1996-2001", *http://www.nist.gov/speech/publications/papersrc/odyssey_paper1.pdf*.

[3] E. S. Parris and M. J. Carey, "Discriminative Phonemes for Speaker Identification", in *Proc. of the 1994 International Conference on Spoken Language Processing (ICSLP 94)* pp. 1843-1846.

[4] R. Auckenthaler, E. S. Parris and M. J. Carey, "Improving a GMM speaker verification system by phonetic weighting", *Proc. of the 1999 IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP 99)*.

[5] M. Newman, L. Gillick, Y. Ito, D. McAllaster and B. Peskin, "Speaker verification through large vocabulary countinous speech recognition", in *Proc. of the 1996 International Conference on Spoken Language Processing (ICSLP 96)*, pp. 2419-2422.

[6] J. Ø. Olsen, "A two-stage procedure for phone based speaker verification", *Pattern Recognition Letters*, vol. 18, pp. 889-897, 1997.

[7] D. Petrovska-Delarcrétaz, J. Černocký, J. Hennebert and G. Chollet "Segmental approach for acoustic speaker verification", *Digital Signal Processing*, vol. 10, pp. 198-212, 2000.

[8] Q. Jin, T. Schultz and A. Waibel, "Phonetic speaker identification", in *Proc. of 7th International Conference on Spoken Language Processing (ICSLP 2002)*, September 16-20, 2002, Denver, Colorado.

[9] D. Gutman and Y. Bistritz, "Speaker verification using phone-adapted gaussian mixture models", in *Proc. of the XI European Signal Processing Conference, (EUSIPCO-2002)*, September 3-6, 2002, Toulouse, France.

[10] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE transactions on Speech and Audio Processing*, vol. 2, pp. 291-298, April 1994.

[11] C. Becchetti and L. R. Prina, *Speech Recognition - Theory and C++ Implementation*, Wiley, 1999.

[12] W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status", in *Proc. DARPA Workshop on Speech Recognition*, Feb. 1986, pp. 93-99.

[13] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A phonetically balanced continuous speech, telephone bandwidth speech database", in *Proc. of the 1990 IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP90)*, pp. 109-112.

Authorized licensed use limited to: TEL AVIV UNIVERSITY. Downloaded on January 20, 2009 at 11:04 from IEEE Xplore. Restrictions apply.