Constrained Cepstral Speaker Recognition Using Matched UBM and JFA Training

Michelle Hewlett Sanchez[1,2], *Luciana Ferrer*[1], *Elizabeth Shriberg*[1], *Andreas Stolcke*[1]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025, U.S.A.
[2]Stanford University, Stanford, CA 94305, U.S.A.
{mhewlett,lferrer,ees,stolcke}@speech.sri.com

Abstract

We study constrained speaker recognition systems, or systems that model standard cepstral features that fall within particular types of speech regions. A question in modeling such systems is whether to constrain universal background model (UBM) training, joint factor analysis (JFA), or both. We explore this question, as well as how to optimize UBM model size, using a corpus of Arabic male speakers. Over a large set of phonetic and prosodic constraints, we find that the performance of a system using constrained JFA and UBM is on average 5.24% better than when using constraint-independent (all frames) JFA and UBM. We find further improvement from optimizing UBM size based on the percentage of frames covered by the constraint.

Index Terms: Speaker Recognition, Cepstral Features, Constraints, Joint Factor Analysis

1. Introduction

One of the most successful approaches to speaker identification models Mel frequency cepstral coefficients (MFCCs) using Gaussian mixture models (GMM) [1] and employs joint factor analysis (JFA) for channel variability compensation [2]. In this and other similar approaches, typically all frames of speech are modeled together.

Some previous research, however, has explored the extraction of cepstral features from only certain regions, to reduce variability from differences in speech content. Sturim et al. constrained a cepstral GMM using a set of frequent words [3]. Baker et al. expanded on this work, constraining on syllables rather than the entire word [4]; Bocklet and Shriberg [5] studied phonetic, syllable-based, and pause-based constraints. Both Park et al. [6] and Shriberg [7] review other studies that condition the regions of cepstral feature extraction on linguistic information such as words and phones.

In this paper, we explore a question for constraint modeling not addressed in earlier work: should the universal background model (UBM) and/or the JFA use all frames, or only the frames in the relevant constraint? We will use the term *constraint-independent* to refer to the use of all frames to estimate the UBM and JFA parameters, and *constraint-dependent* to refer to the use of only the frames within the constraint of interest. For constraint-dependent modeling, we further ask whether and how to optimize UBM size given that constraints differ in sparsity.

Speaker verification experiments are performed on a database of Arabic male speakers that contains a large number of sessions for each target speaker.

2. Constrained cepstral features

We explore a range of constraint types. Unit-based constraints are regions constrained by specific syllable, phone, or subphone regions. Prosodic or acoustic constraints are regions constrained by voicing, energy and pitch values, or by pitch and energy slopes. Some examples are regions including voiced frames or regions constrained by the sign of the slope in the energy and the pitch. Turn-taking or discourse-related constraints are regions constrained by their location relative to spurts (regions of speech without long pauses). Examples include regions of N frames at the beginning and/or end of a spurt. Speaking rate constraints are regions constrained by different measures of speaking rate. Some examples are regions with a min/max value of the phone length normalized by phone and speaker statistics or windows with a min/max number of phones per unit time where the unit is the window length.

3. Modeling approach

Features are based on 60 MFCCs consisting of 20 coefficients with cepstral mean subtraction in addition to first and second derivatives. The features are used to model both the targets and impostors using GMMs. We trained a background GMM with held-out data to use as our UBM. We tested two different methods of speaker adaptation. First, we experimented with MAP adaptation, in which the GMM is adapted to each speaker's features using a maximum a posteriori estimation of the means. The second methodology used was JFA. JFA assumes the means of the speaker's model are given by **m'=m+Vy+Ux** where **m** is the background supervector (concatenation of the GMM means), **V** is a rectangular low-rank matrix in which columns are the directions of speaker variability known as eigenvoices (ev), and **U** is another rectangular low-rank matrix in which columns are the number of directions of channel variability, or eigenchannels (ec). Values **x** and **y** are learned from the sample. The score obtained for each test sample is the estimated likelihood ratio between the speaker's model and the UBM.

The modeling approach for the constrained cepstral features is the standard JFA method normally used for modeling all frames, except that statistics are now computed using features extracted only from regions of interest. We discuss in this paper two different ways to obtain the parameters of the models: constraint-independent modeling where UBM and JFA matrices are obtained using all frames (baseline case) and constraint-dependent modeling where UBM and JFA matrices are obtained using only frames over the constrained regions of interest.

The optimal size of the UBM (number of Gaussians) for each constraint was investigated empirically, since this could be affected both by frame sparsity and by inherent homogeneity of

the constraint. We will use the term *frequency* to refer to the percentage of frames covered by the constraint. This measure is computed with respect to all speech (nonpause) frames; the baseline in this case would be 100% or all frames of speech. Note that the frequency measure is a function of both (1) the frequency of the occurrence of the constraint, and (2) the constraint length.

The results discussed in this paper involve each constraint performance by itself not in combination with the all-frame baseline. In this paper, we explore each constraint on its own since our goal is to understand the UBM and JFA parameters and how they are affected by the frequency of the constraint. Our end goal is to investigate the performance of one or more constraints in combination with the baseline.

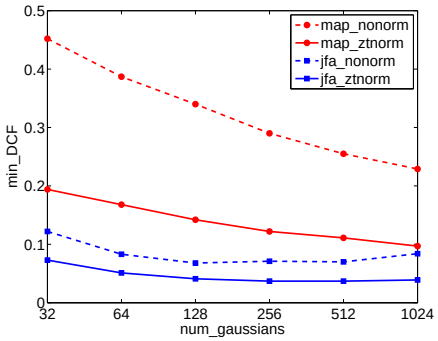


Figure 1: *Performance of baseline system.*

4. Datasets

The male Arabic database is composed of data from several sources. IntelCenter (IC) provided both audio and video samples of 57 Arabic male speakers with as many as 51 sessions each. The GALE Broadcast News (BN) data, contains 2124 male speakers with as many as 80 sessions each [8]. These two datasets were used because of the large number of sessions for each speaker. Since the GALE database was not originally designed for speaker recognition, we further processed the data to assure the accuracy of target speaker labels for our experiments. The Mixer corpus as provided by NIST for the 2004 and 2005 Speaker Recognition Evaluations (SRE04 and SRE05) and LDC data in various Arabic dialects were also used. Together, the NIST and LDC corpora provided 135 male speakers with a total of 257 sessions. We limit this study to male speakers, both because of data availability, and because state-of-the-art systems use gender-dependent modeling techniques that would have added complexity to the experiments.

The target speakers were chosen from the IC and GALE BN databases. There were eight target speakers from IC that had at least 14 sessions each, and 35 target speakers from GALE that had at least 16 sessions each and for which we were confident of the speaker identity. The remaining speakers not used as targets were split for use as impostors and for UBM and JFA parameter training. The UBM used only one or two sessions for each of 1224 speakers, while for JFA training we chose 142 of those speakers which had at least 6 sessions each. The remaining speakers were used to create impostor samples against the target speakers, using 525 speakers for development and 524 for evaluation. The Z-normalization and T-normalization matrices (ZTNORM) were selected from the same speakers used for the UBM, but selecting only one session for each speaker. The GALE data made up 90% of the data used to train the UBM, JFA, and ZTNORM. The remaining 10% came from the small amount of Arabic LDC and NIST data.

5. Experiments

5.1. Task setup

We created five models for each of the target speakers using approximately 120 seconds from each of three different sessions. Test samples are given by 30-second snippets randomly selected from the available target and impostor sessions. We chose not to use all snippets from each session to avoid having highly correlated samples that might lead to biased results. We tested each of the models against all other available sessions for the particular target speaker. We also tested each of the models against all the impostor samples and other target speakers. For each model in this task, this results in an average of 20 target samples and 1700 impostor samples.

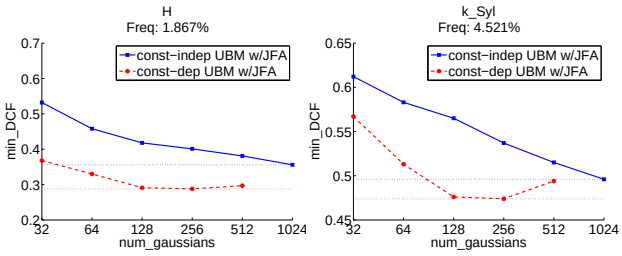
5.2. Baseline all-frame system

We performed a large number of calibration experiments including comparing MAP to JFA, comparing using no normalization to ZT-normalization (Z-normalization followed by T-normalization), and varying the UBM size. The goal of these experiments was to obtain the best possible baseline for comparison purposes. Figure 1 shows our calibration experiments on the baseline system. We use the traditional NIST minimum detection cost function (DCF) as our performance measure ($9.9\times\mathbf{P}_{\mathbf{fa}}+\mathbf{P}_{\mathbf{miss}}$, $\mathbf{P}_{\mathbf{fa}}$ is false alarm probability, $\mathbf{P}_{\mathbf{miss}}$ is miss probability). We chose this performance measure instead of equal error rate (EER) because it operates at a higher miss rate, which gives more target errors since the number of target samples is much smaller than the number of impostor samples.

Both JFA and ZTNORM give large gains. Experiments were performed on the baseline system to find the optimal JFA parameters: 100 eigenchannels (ec) and 100 eigenvoices (ev). The performance of our system was optimal using JFA and ZT-NORM for a UBM size of 512. The corresponding EERs range from 7.6% to 0.93%. Based on SRI's performance on NIST SRE10 [9], we believe that this baseline is close to the state of the art for cepstral systems.

5.3. Constraint-independent systems

A constraint-independent system uses all frames (as in the baseline system) to train the UBM. This method was tried with both MAP and constraint-independent JFA, which also uses all frames to train the JFA matrices, with and without ZT-NORM. Our hypothesis was that if one trains the UBM with all frames, then the UBM size will need to be large to account for a constraint with a small frequency. We obtained the constraint-independent results for 104 different constraints ranging in frequencies from 0.17% to 38.52%. We ran experiments on all of the constraints for six different sizes of the constraint-independent UBM: 32, 64, 128, 256, 512, and 1024 Gaussians. Over all constraints, the optimal performance was found using UBM size of 512 with constraint-independent JFA and ZTNORM, as it was for the baseline. We will use this as our baseline system for each constraint. We did not try 2048 Gaussians because models with even larger number of Gaussians are often not practical due to the time it takes to train and evaluate them.



(a) (b)

Figure 2: *Constraint-independent versus constraint-dependent results for* (a) *Arabic phone* H *(unvoiced pharyngeal fricative) and* (b) *all Arabic syllables with* k *phone (unvoiced velar stop).*

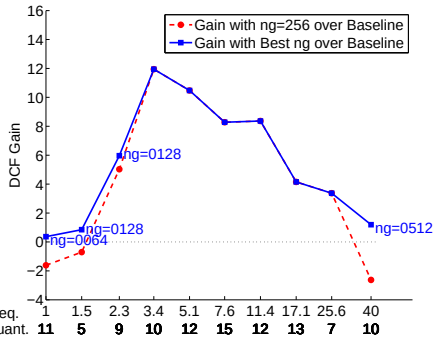


Figure 3: *Average gain for constraint-dependent system with JFA used for analysis.* Freq. *is maximum frequency in each bin and* Quant. *is number of constraints in each bin.* ng=xxxx *shows the optimum number of Gaussians used for the UBM when this optimum was not* 0256.

5.4. Constraint-dependent systems

A constraint-dependent system uses only the frames selected by a given constraint to train the UBM. This method was tried with both MAP and constraint-dependent JFA, which likewise uses only selected frames in training the JFA matrices, with and without ZTNORM. Our hypothesis was that if one trains constraint-specific UBMs, then the model size suitable for representing the infrequent constraint should be smaller than when the UBM is trained on all frames. We obtained the constraint-dependent results for the same 104 constraints as before. We ran experiments for five different sizes of the constraint-dependent UBM: 32, 64, 128, 256, and 512 Gaussians. We did not try 1024 Gaussians for two reasons. First, since the constraint-independent systems were optimal with a size of 512 Gaussians, we would not expect the optimal constraint-dependent UBM to exceed this size. Second, most of the constraints show degradations for the constraint-dependent system when the UBM size reaches 512.

Figure 2 shows constraint-independent results and constraint-dependent results for various UBM sizes, for two of the constraints. In all cases, JFA was used and ZTNORM is applied to scores. The name and frequency of each constraint are given in the figure. The dotted horizontal lines show the best DCF achieved using the constraint-independent UBM. Both phone *H* (unvoiced pharyngeal fricative) and all syllables containing *k* phone (unvoiced velar stop) show significant gains when using the constraint-dependent method. These cases are representative of most other constraints, except for very infrequent constraints for which the constraint-dependent model is not robust. Furthermore, most of the very frequent constraints do not show significant gains over the baseline system since, for frequent constraints, the resulting models behave much like the constraint-independent baseline system, which uses all frames. The most significant gains from the constraint-dependent approach are obtained in the frequency range between 2% and 20%.

On average over all constraints, the optimal performance was found using the constraint-dependent approach with a UBM of size 256. This performance is 5.24% better than the average performance of the constraint-independent approach with optimal UBM size of 512. These results support the original idea that the size needed for the UBM for a sparse constraint with constraint-dependent modeling is smaller than that needed with constraint-independent modeling.

We also tried using a constraint-independent UBM and constraint-dependent JFA matrix for each constraint. This approach did not perform as well as constraining both the UBM and the JFA, so we did not investigate it further.

5.5. Analysis

When using a constraint-dependent UBM model of size 256 with JFA, some constraints degraded with respect to using a constraint-independent UBM of size 512. Therefore, we investigated ways to vary the number of Gaussians depending on the frequency of the constraint. We split or binned the constraints based on their frequencies on a log scale. The dotted line in Figure 3 shows the gain for the 10 different frequency bins, over the baseline using a constraint-dependent UBM of size 256. Each tick on the x-axis shows the maximum value of the frequencies that bin holds (Freq.). For example, 1 represents frequencies 0-1%, 1.5 represents frequencies 1-1.5%, and so forth. The number of constraints in each bin (Quant.) is labeled below the frequencies. The figure shows degradations for very low and very high frequencies.

To see whether degradations could be avoided by choosing different UBM sizes based on the constraint frequency, we chose the best size of the constraint-dependent UBM for each bin of frequencies and compared its performance with that of the fixed 256-Gaussian constraint-dependent UBM. The solid line in Figure 3 shows the gain over the baseline when choosing the optimal size of the constraint-dependent UBM for each bin. For constraints with frequencies less than 1.0%, choose a UBM size of 64 Gaussians; for frequencies between 1.0% and 2.3%, use 128; for frequencies greater than 25.6%, use 512. Otherwise, choosing a UBM size of 256 is optimal. This set of UBM sizes obtained as a function of the frequency always achieves positive gains over the baseline.

5.6. Optimizing UBM size for constraint-dependent systems

The described method to find the optimal UBM size for each constraint gives optimistic results since decisions are made based on all constraints within a frequency bin and applied to those same constraints. Consequently, we propose a fair method for finding the UBM sizes as a function of the frequency.

The 104 constraints were randomly split equally into two groups so that the information obtained for group 1 could be used on group 2 and vice versa. We inspected scatter plots for each group of constraints with the gains over the baseline versus the frequency on a log scale for each of the five different UBM sizes. These scatter plots have a somewhat quadratic shape. Hence, for each UBM size and each group of constraints, we calculate quadratic regression curves. Figure 4 shows the five

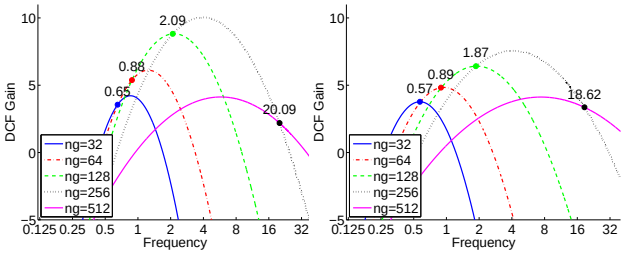


Figure 4: *Regression curves for constraints in* (a) *group 1 and* (b) *group 2 labeled with points of intersection.*

quadratic regression curves for the two groups of constraints.

The points of intersection of the five curves for group 1 are used as the thresholds to choose the optimal UBM size for group 2. For group 2, we chose a UBM size of 32 for constraints with frequencies less than 0.65%, 64 for constraints between 0.65% and 0.88%, and so forth. Similarly, the points of intersection of the five curves for group 2 were used as the thresholds to choose the optimal UBM size for group 1. The thresholds were quite similar across groups; only 5 of the 104 constraints fell between the two groups' thresholds.

Figure 5 shows the same average gain results as Figure 3, except that each tick on the x-axis is the value of the constraint-dependent UBM size (ng) that is optimal based on this thresholding method. The number of constraints for each optimal UBM size (Quant.) is labeled below the frequencies. The dotted line in Figure 5 shows the gain over the baseline when doing constraint-dependent modeling of the UBM with size 256. The solid line in Figure 5 shows the gain over the baseline when choosing the optimal size of the constraint-dependent UBM. A total of 37 constraints out of the 104 did not use a UBM size of 256 with this thresholding method. The average gain in choosing the UBM based on frequency compared to choosing size 256 is 1.57% over these 37 constraints.

Using the proposed frequency thresholding method for choosing constraint-dependent UBM size, the average gain over the baseline over all the constraints is 5.83%. Some constraints had as much as a 24% gain over the baseline when using the obtained UBM size like phone *H* in Figure 2. Syllables containing the *k* phone achieve an 8% gain. Even after choosing the optimal UBM size in the described way, some constraints showed degradation with respect to using the constraint-independent UBM and JFA. The worst degradation was 15% relative.

It is possible that these constraints are, in fact, not so specific in terms of the location of the features that satisfy the constraint in the feature space and, hence, are described fine by the overall distribution (over all frames). On the other hand, some constraints are very specific and do benefit from constraint-dependent modeling.

6. Conclusions

We model constrained cepstral features using the JFA technique and compare two approaches for training the UBM and JFA parameters: a constraint-independent approach and a constraint-dependent approach. We show that the constraint-dependent approach outperforms the constraint-independent one on average over all the constraints. Furthermore, we find that a smaller number of Gaussians is needed in the UBM when it is trained using only constraint-specific frames than when trained on all frames. Finally, we find that a simple method for predicting the

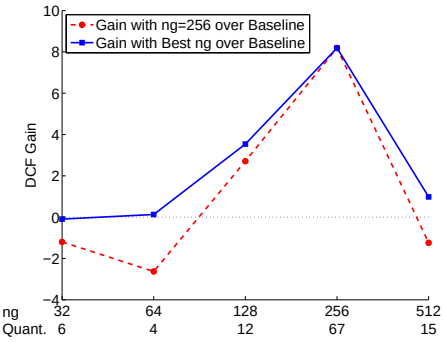


Figure 5: *Average gain for constraint-dependent system with JFA using quadratic regression thresholds.* ng *is number of Gaussians used for UBM using thresholds and* Quant. *is number of constraints for each optimal UBM size.*

size of the constraint-dependent UBM outperforms the systems obtained with a fixed UBM size on average over all constraints.

We believe a constrained UBM and JFA performs better than an unconstrained UBM and JFA because matching the content of the frames is more important than using a larger number of frames that contain unmatched content.

7. Acknowledgments

We thank Nicolas Scheffer for the development of the JFA code used for the experiments in this paper and his guidance on different issues. We thank Martin Graciarena and Professor Robert M. Gray for their valuable discussions. This research effort was funded under DISA Encore II contract HC1028-08-D-2027 with Unisys Corporation serving as the prime on behalf of the Department of Defense.

8. References

[1] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[3] D. Sturim, D. Reynolds, R. Dunn, and T. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. 677–680.

[4] B. Baker, R. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllable events for text-independent speaker verification," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 2429–2432.

[5] T. Bocklet and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009.

[6] A. Park and T. Hazen, "ASR dependent techniques for speaker identification," in *International Conference on Spoken Language Processing*, 2002, pp. 1337–1340.

[7] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller, Ed., no. 4343 in *Lecture Notes on Artifical Intelligence*. Springer, 2007, pp. 241–259.

[8] J. Zheng, W. Wang, and N. Ayan, "Development of SRI's translation systems for broadcast news and broadcast conversations," in *Interspeech*, Brisbane, Australia, 2008.

[9] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011.