ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Cluster-dependent Modeling and Confidence Measure Processing for In-Set/Out-of-Set Speaker Identification

Pongtep Angkititrakul, John H.L. Hansen, Sepideh Baghaii

Robust Speech Processing Group, Center for Spoken Language Research (CSLR)
University of Colorado at Boulder, Boulder, CO 80302, USA
{angkitit,jhlh,baghaii}@cslr.colorado.edu, web: http://cslr.colorado.edu

10.21437/Interspeech.2004-537

Abstract

In this paper, we propose an approach to address the problem of text-independent open-set speaker identification. The in-set speakers are clustered into smaller subsets without merging speaker models. The Anti-Speaker or Background Model is then adapted for each subset which minimizes the identification errors of the pseudo impostors during the training stage. Score normalization is applied to align all the in-set speaker score distributions to share a single scale. Finally, confidence measure processing is used to identify in-set versus out-of-set speakers. Experiments with TIMIT and the CU-Accent corpora show an improvement in Equal Error Rate on the average of 20.28% and 8.35% over the baseline performance respectively. Finally, a probe experiment is also included that considers prosody for in-set speaker detection.

1. Introduction

The ability to ensure security and proper access to private information of a specific group of people (i.e., communication networks, bank sharing-accounts, health-care centers, personal device assistants, location access) is a general goal for speaker identification/verification. The objective of Open-set Speaker Identification (OSI) is to make a decision regarding an input speaker as being a legitimate member of an enrolled speaker set or not. The conventional OSI framework is first to *identify* a claimant as the most likely in-set speaker, and then further *verify* the claimant as whether to accept or reject s/he as the hypothesized speaker.

Some prior studies of OSI which employed Gaussian Mixture Model(GMM) as the underlying modeling technique [10, 6, 4] together with expanded novel processing have shown promising results. The problem of Open-set Speaker Identification is a challenging research problem due to relaxing the *closed-world assumption* (i.e., as in closed-set speaker identification). Ideally, we would prefer that the in-set speakers share some unique speaker trait which can easily distinguish their voices from impostors (e.g., gender, age, regional accent, occupation for word selection, etc). In practice, there are distinct dimensions of variability which greatly affect the performance of speaker recognition systems such as inter-speaker variations at both the segmental and suprasegmental levels [3], and communication channel effects (i.e, telephone handset, background noises) [5]. Our primary studies have shown that the system performance tends to degrade–in terms of Equal Error Rate (EER)–as the size of the in-set speakers increases [1]. In this paper, we split the in-set speakers into individual smaller tree based subsets using pitch information. We conjecture that some in-set speaker characteristics are more related than others (or not at all). Instead of performing the hypothesis testing with the conventional Likelihood Ratio Test (LRT) of the most likely in-set speaker against the Universal Background Model (UBM) or Anti-Speaker model, we test the hypothesis of the most likely in-set speaker versus the corresponding Cluster-dependent Background Model (CBM). The use of a Cluster-dependent Background Model is motivated by the idea to shift the *random pooled* Universal Background Model with a better discriminate cluster of speakers from the other speakers, without losing the in-set speaker discrimination capability. Minimum Classification Error criterion is investigated in this work for adapting each subset background model. Score normalization (Zero normalization) is applied to align all in-set speaker score distributions to share a single scale. Confidence Measure (CM) is computed and used as the final comparative score. System performance with different error matrices are also investigated.

This paper is organized as follows: First, we briefly provided an overview of the OSI and the baseline system. In Sec. 3, we introduce the general framework for background model adaptation using a minimum classification error paradigm. Next, we discuss our score normalization and confidence measure techniques. Experiments and results are presented in Sec. 5. In Sec. 6, we discuss our probe study using pitch contours, with conclusions in Sec. 7.

2. Baseline System

A state-of-the-art GMM-UBM with the conventional Likelihood Ratio Test (LRT) is used as our baseline system. A UBM is trained from non-target speakers from a development set using the Expectation Maximization (EM) algorithm. The probability density function (*pdf*) is modeled by an M-component GMM, $\Lambda_{UBM} : \{\omega_0, \mu_0, \Sigma_0\}$. For each in-set speaker, a speaker-dependent GMM $(\Lambda_n : \{\omega_n, \mu_n, \Sigma_n\}, 1 \le n \le N)$ is created by Maximum A Priori (MAP) adaptation of the UBM parameters [9]. For our experiments, only mean adaptation is applied. Let $\mathbf{X}_\tau = \{X_{\tau 1}, \ldots, X_{\tau T}\}$ denote the observation sequence of length T extracted from the τ-th utterance. The average log-likelihood (LLH) score generated by each GMM is computed as,

$$\mathcal{L}(\mathbf{X}_\tau|\Lambda_n) = \frac{1}{T}\sum_{t=1}^{T}\log\sum_{i=1}^{M}\omega_{ni}\mathcal{G}_{ni}(X_{\tau t}), \qquad (1)$$

where ω_{ni} is the weight of the i-th component, and $\mathcal{G}_{ni}$ is the multivariate Gaussian probability density function with mean μ_{ni} and covariance matrix Σ_{ni}, which is assumed diagonal. The hypothesized in-set speaker, κ_τ, is the most likely speaker whose GMM generates the highest LLH score:

$$\kappa_\tau = \arg\max_{1\le n\le N}\mathcal{L}(\mathbf{X}_\tau|\Lambda_n). \qquad (2)$$

Next, the LLH score of the hypothesized speaker is compared against the LLH score generated from the Anti-Speaker GMM,

This work was supported by the U.S. Air Force Research Laboratory, Rome NY, under contract F30602-03-1-0110.

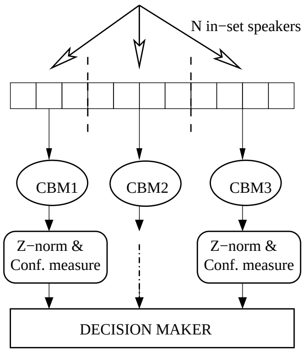

Figure 1: *Subset identification framework.*

Λ_{anti}, (i.e., Λ_{UBM} for baseline system):

$$\mathcal{D}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}) = \mathcal{L}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}) - \mathcal{L}(\mathbf{X}_\tau|\Lambda_{UBM}). \qquad (3)$$

If this difference or relative LLH score is greater than a predefined threshold, we accept the observation sequence as belonging to the hypothesized speaker(in-set); otherwise, we reject it as an impostor(out-of-set).

3. Cluster-dependent Background Model

Fig. 1 shows our subset identification framework. After clustering speakers into individual subsets, the UBM is adapted to minimize classification errors of the pseudo impostors to each subset, namely Cluster-dependent Background Model (CBM). It is important to note that our clustering criterion is for grouping speakers into subsets without merging any speaker models.

A Minimum Classification Error (MCE) training paradigm has previously been applied successfully for text-independent speaker identification [10] and automatic speech recognition [7]. For our formulation, we employ a similar scheme to adapt the UBM as follows: The difference LLH score is now defined as:

$$\mathcal{D}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}) = \mathcal{L}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}) - \mathcal{L}(\mathbf{X}_\tau|\Lambda_{CBM_{\kappa\tau}}), \qquad (4)$$

where $\mathcal{L}(\mathbf{X}_\tau|\Lambda_{CBM_{\kappa\tau}})$ is the LLH score of a CBM corresponding to the hypothesized speaker κ_τ. Each LLH score is embedded into a smooth empirical *loss function* which approximates the loss directly related to the number of identification errors:

$$l_{CBM}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}) = \frac{1}{1 + exp(-\alpha\mathcal{D}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}))}. \qquad (5)$$

The positive constant α is used to control the slope of the decision threshold. Given the set of training observations $\mathbf{X}$ from pseudo impostors, the total loss is found as:

$$l_{CBM}(\mathbf{X}) = \sum_\tau l_{CBM}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}). \qquad (6)$$

Using a Gradient Probabilistic Descent(GPD) algorithm, it is possible to achieve a local minimum of the loss function using an iterative procedure based on individual token loss. The parameter update formula is:

$$\Lambda^{\tau+1}_{CBM} = \Lambda^{\tau}_{CBM} - \varepsilon_\tau \nabla_{\Lambda^{\tau}_{CBM}} l_{CBM}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}), \qquad (7)$$

where $\nabla_{\Lambda^{\tau}_{CBM}} l_{CBM}(.)$ is the gradient of the loss function, and ε_τ is the learning step size. The model parameters are updated on a training token by training token basis(i.e., every τ-th utterance). The seed model for each CBM is the UBM. Also, only mean parameters are adapted for our experiments.

4. Score Normalization & Confidence Measure

4.1. Zero Normalization (Z-norm)

Z-norm is a normalization technique which uses mean and standard deviation estimation for distribution scaling [2]. The advantage of Z-norm is that the estimation of the normalization parameters can be performed off-line during training. A speaker model is tested against pseudo-impostor utterances and the different LLH scores are used to estimate a speaker specific mean and standard deviation for the impostor distribution. The normalization has the form,

$$\mathcal{Z}_\tau = \frac{\mathcal{D}(\mathbf{X}_\tau|\Lambda_{\kappa\tau}) - \mu_{\kappa\tau}}{\sigma_{\kappa\tau}}, \qquad (8)$$

where $\mu_{\kappa\tau}$ and $\sigma_{\kappa\tau}$ are the estimated impostor mean and standard deviation for speaker model κ_τ, and $\mathcal{Z}_\tau$ is the new normalized difference LLH score (e.g., from Eq. 4 and 8). Z-norm aligns all in-set speaker distributions to share a single scale, and then allows us to use a single threshold across all hypothesized speakers.

4.2. Confidence Measure (CM)

Another useful design specification is the percentage of acceptable utterances [8]. A Confidence Measure of the test utterance τ can be represented as:

$$C_\tau = \frac{\mathcal{Z}_\tau}{|\mathcal{L}(\mathbf{X}_\tau|\Lambda_{\kappa\tau})|}. \qquad (9)$$

Confidence Measure is a preferable statistic due to its stability and limited numerical range. One benefit of combining the Z-norm and Confidence measure is to help better control the threshold for desired False-acceptance(FA) rate and confidence of acceptance.

5. Experiments

5.1. Front-End Processing

The speech analysis frame rate is set to 30 ms with a 10 ms skip rate. All speech is pre-emphasized with the filter $(1 - 0.95z^{-1})$, and a Nineteen-dimensional Mel-Frequency Cepstral Coefficient (MFCC) vector is extracted per frame. Silence and low-energy speech parts are removed using a general energy detection technique. Cepstral Mean Normalization(CMN) is applied to each utterance to reduce channel based spectral shaping.

5.2. Experimental Setup

5.2.1. TIMIT

A set of 60 male speakers was randomly selected as the in-set speakers. Four particular sizes of in-set speakers from these speakers are considered (e.g., four sets of 15 speakers, two sets of 30 speakers, two overlapped sets of 45 speakers, and one set of all 60 speakers). Another set of 120 male speakers was randomly selected as a development set (60 speakers were used to train the UBM, another 60 speakers were used as the pseudo impostors). The remaining 258 TIMIT male speakers were used as the impostors during the testing stage. The “si” and “sx” sentences are used during training and developing stages, while the “sa” sentences are set aside for testing.

The selection of speakers from TIMIT for in-set was essentially random with no distinct in-set characteristic. In the second in-set evaluation, we consider in-set speaker which have a distinguishing characteristic.

5.2.2. CU-Accent

CU-Accent corpus [11] was collected across a telephone channel, digitized at 8000 Hz and stored in 16-bit linear PCM format. The

Table 1: *Comparison EER of in-set versus out-of-set identification. “CBM” indicates the difference LLH of the hypothesized speaker and corresponding CBM. “CBM&Znorm” is the difference LLH with zero normalization. “CBM&Znorm&CM” is the confidence measure of the difference LLH with zero normalization (Native Speakers: En=English, Tu=Turkish, Th=Thai, Fr=French, Ma=Mandarin).*

	TIMIT				CU-Accent				
# in-set speakers	15	30	45	60	9(En)	10(Tu)	12(Th)	14(Fr)	28(Ma)
# out-of-set speakers	258	258	258	258	64	63	61	59	45
					(Tu,Th,Fr,Ma)	(En,Th,Fr,MA)	(En,Tu,Fr,Ma)	(En,Tu,Th,Ma)	(En,Tu,Th,Fr)
Baseline	10.38	12.89	14.60	16.21	18.91	19.13	19.30	22.80	26.70
CBM	–	11.65	13.15	14.57	18.55	19.22	18.92	22.58	27.01
CBM&Znorm	–	11.74	11.31	13.64	17.98	18.34	20.47	23.13	26.71
CBM&Znorm&CM	–	10.58	11.08	13.16	17.55	16.47	18.48	20.15	25.42

database contains speakers with a variety of foreign accents (i.e., Mandarin, Thai, French, Turkish, and others) with multiple calling sessions. In this study, we selected 73 speakers who have at least 3 calling sessions as a speaker sample space (i.e., 9 American[4m/5f], 14 French[6m/8f], 28 Mandarin[9m/18f], 12 Thai[9m/3f], and 10 Turkish[3m/7f]). While speakers from one accent class play the roles of in-set speakers, the remaining speakers from other accent classes will act as the impostors during testing. One session of the remaining 110 speakers in the database was used as a development set (58 speakers were used to train the UBM, and 52 speakers were used as the pseudo impostors). The isolated-word section, where each speaker spoke 5 tokens of 23 words per session, was used in our experiments (each word is between 620–1150 ms. in duration.) The in-set speaker models were trained from sessions 1 and 2 (approximately 1.5 min. worth of speech for each speaker). Session 3 was used for evaluation (7528 testing words for each experiment).

5.3. Evaluation Results

A UBM is trained from speakers in the development set, with 64 Gaussian components. A single speaker-dependent GMM for each in-set speaker is then estimated from the UBM based on MAP adaptation. The number of Gaussian components is also fixed to 64 for all speakers in our experiments. The baseline system employs the difference LLH from Eq.(2) and (3) for hypothesis testing. In this work, we are interested in the correct identification of in- versus out-of-set. We consider pseudo-false acceptance (accepting as an incorrect in-set speaker) as a correct identification.

For our cluster-based scheme, we used the Snack Sound Toolkit [12] to find the mean pitch of each in-set speaker. Speakers within the same pitch range are grouped together. For TIMIT, the fixed size of 15 speakers per subset is considered (i.e., two subsets for 30 in-set size, three subsets for 45 in-set size, and four subsets for 60 in-set size). Each CBM is adapted from the UBM, using in-set speaker GMMs from a subset and pseudo impostors from the development set. For the CU-Accent database, a fixed size of 4 (or 5) speakers per subset is considered (i.e., two subsets for En9, two subsets for Tu10, three subsets for Th12, three subsets for Fr14, and six subsets for Ma28). The α is set to 1.0 and 0.1 for TIMIT and CU-Accent respectively. The initial learning rates are 0.01 (TIMIT) and 0.004 (CU-Accent) with a half-rate scaling down every five iterations. Fifteen adapting iterations are applied to each CBM.

5.3.1. *Equal Error Rate*

In this experiment, the most-likely in-set speaker is identified (Eq.2) and three algorithm steps (CBM, CBM+Znorm, CBM+Znorm+CM) for LLH score generation. Table 1 shows the identification Equal Error Rate (EER) for the TIMIT and CU-Accent Corpora at different in-set sizes. From Table 1, we see a consistent increase in EER as the in-set speaker size increases for both databases. We see a small gain from using CBM for CU-Accent and a higher increase for TIMIT experiments. Consistent improvement is achieved after Z-norm and Confidence Measure processing. Our assumption is that the experimental setup for CU-Accent grouping in-set speakers based on their native languages which already shared some common traits among the in-set speakers from the out-of-set speakers. Also, the CU-Accent corpus contains matched and mismatched telephone handsets between training and testing which degrades the identification performances [5]. Again, the evaluation with TIMIT is based on a sentence utterance, while CU-Accent is based on an isolated-word utterance. Fig. 2 shows example DET curves of TIMIT with 45 in-set speakers and CU-Accent with 14(Fr) in-set speakers. We see measurable improvement for both corpora using CBM with Z-norm and Confidence Measure vs. baseline performance.

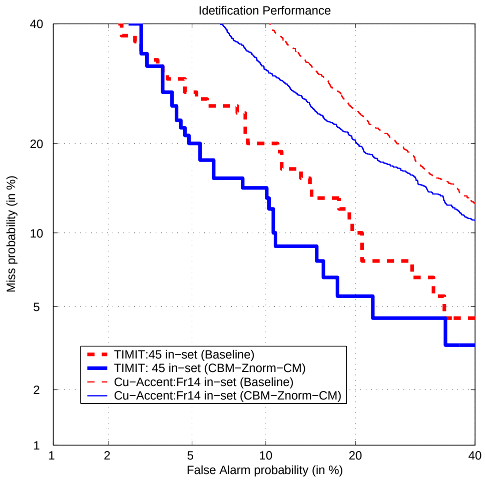

Figure 2: *DET curve illustrates the system performance.*

5.3.2. *False Acceptance & False Rejection*

To further study the relationship of system performance compared to its subset performance, we consider the following experiment. Suppose for each subset, we set a confidence threshold so that the False Alarm(FA) rate is close to the False Rejection(FR) rate. At the testing stage, we accept a claimant as an in-set speaker if the confidence measure is greater than a pre-defined threshold for each corresponding subset. Table 2 compares FA/FR rates of the identification performance with its subsets. From this table, the FA/FR rate of the in-set has the following relationship as:

$$\mathcal{E}_r(\mathcal{M}) \approx \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \epsilon_r(m), \mathcal{E}_a(\mathcal{M}) \approx 1 - \prod_{m=1}^{\mathcal{M}} (1 - \epsilon_a(m)), \quad (10)$$

Table 2: *Comparison FA and FR rates (%) of the in-set performance to its component subsets, with the pre-defined threshold for each subset.*

Experiments	FA/FR (%)				
	subset1	subset2	subset3	subset4	overall
TIMIT: 30 in-sets	7.14/6.67	4.58/3.33	∅	∅	11.05/5.0
TIMIT: 45 in-sets	7.81/6.67	4.86/6.67	6.25/6.67	∅	15.89/6.67
TIMIT: 60 in-sets	8.42/6.67	5.45/6.67	9.74/10.0	2.46/0.0	20.74/5.83
CU-Accent: 9(En)	15.55/16.46	6.88/8.08	∅	∅	22.69/12.90
CU-Accent: 12(Th)	15.09/12.83	10.52/10.04	10.22/8.17	∅	26.81/9.68
CU-Accent: 14(Fr)	11.92/7.17	23.51/26.87	7.70/7.91	∅	33.89/12.60

where $\mathcal{M}$ is the number of subsets, $\mathcal{E}_r(\mathcal{M})$ and $\mathcal{E}_a(\mathcal{M})$ are FA and FR probabilities for $\mathcal{M}$ subsets respectively, and $\epsilon_r(m)$ and $\epsilon_a(m)$ are FR and FA probabilities of the m-th subset respectively. Such a decision criterion will increase FA as the number of subsets increases, while the FR is the average of all subsets. This experiment may help us to predict the system performance when additional in-set speakers are added to the system in the future. From this table, we see that FA/FR performance degrades as the number of in-set speakers increases. We also see that the FA and FR rates are comparable.

6. Prosody Probe Study

We would like to close the paper with our prosody probe study for OSI. For development of cluster-dependent models, we used mean pitch to establish the tree for in-set speaker subgroups. We feel that it is also possible to use more extensive prosody structure to partition and assist in the formulation of in-set speaker space. Here, we consider overall pitch contour structure as a discriminating. The experiment was to determine if in-set speakers share a common trait such as native language, is there a common prosodic trait that can help group in-set speakers? Using the CU-Accent corpus, nineteen speakers of native American English(AE) are selected as in-set speakers, and 32 speakers who are not native AE are selected as out-of-set speakers. The gray scale histogram of pitch contours of the same sentence from all speakers are shown in Fig. 3. For each speaker, the normalized pitch contour is used to reduce speaker-dependent traits (i.e., we subtract mean pitch and build contour histograms). We can see that the mean slope of the pitch contour of in-set speakers is smaller than the out-of-set speakers. We believe that such a common trait could be useful for improving OSI performance, and our future work will consider integrating this knowledge into our system.

7. Conclusions and Future Work

In this paper we investigated the problem of In-Set/Out-of-Set Speaker Identification by splitting in-set speakers into smaller subsets and adapting a UBM model to better discriminate each speaker subset from the pseudo impostors. Score normalization and confidence measure processing were applied to align all the score distributions to share the same scale. The evaluations showed consistent improvement in Equal Error Rate for all experiments. We also showed the relationship of False Acceptance and False Rejection rates of the system compared to its subsets. The proposed framework showed promising results as in-set speaker size was increased. We believe that further study on clustering in-set speakers and anti-speaker model development will further improve system performance. Our future work will consider such issues and also hierarchical hypothesis testing.

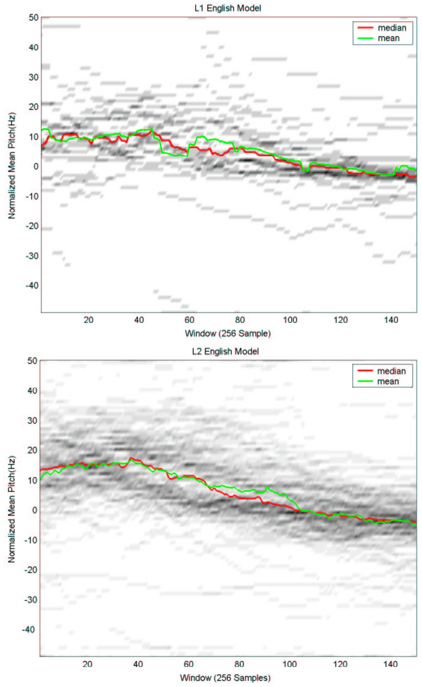

Figure 3: *Gray scale histograms of pitch contour of in-set speakers (upper) and out-of-set speakers (lower).*

8. References

[1] Angkititrakul, P. and Hansen, J.H.L., “Identifying in-set and out-of-set speakers using neighborhood information,” in *ICASSP 2004* (accepted and to appear).

[2] Auckenthaler, R., Carey, M., and Lloyd-Thomas, H., “Score Normalization for Text-Independent Speaker Verification Systems,” *Digital Signal Processing.* **10**(2000), 42–54.

[3] Doddington, G., “Speaker Recognition Evaluation and Methodology: An Overview and Perspective,” Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 60–66, 1998.

[4] Gong, Y., “Noise-robust open-set speaker recognition using noise dependent Gaussian mixture classifier,” in *Proc. ICASSP 2002*, pp. I.133–136.

[5] Heck, L., and Weintraub, M., “Handset-dependent background models for robust text-independent speaker recognition,” in *ICASSP 1997*, pp. 1071–1074.

[6] Jiang, H., and Deng L., “A Bayesian Approach to the Verification Problem: Application to Speaker Verification,” IEEE Trans. Speech and Audio Proc., 9(8):874–884, Nov. 2001.

[7] Juang, B.-H., Chou, W., and Lee, C.-H.,, “Minimum classification error rate methods for speech recognition,” IEEE Trans. Speech and Audio Proc., 5(3):257–265, 1997.

[8] Li, Q., and Juang, B.-H., “Speaker Authentication,” *Pattern Recognition in Speech and Language Processing* , Elec. Eng. & Appl. Sig. Proc. series(**12**), CRC Press, 2003.

[9] Reynolds, D., Quatieri, T., and Dunn, R., “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing.* **10**(2000), 19–41.

[10] Siohan, O., Rosenberg, A.E., and Parthasarathy, S., “Speaker identification using minimum classification error training,” in *Proc. ICASSP 1998*, pp. 109–112.

[11] http://cslr.colorado.edu/accent

[12] http://www.speech.kth.se/snack/