IDENTIFYING IN-SET AND OUT-OF-SET SPEAKERS USING NEIGHBORHOOD INFORMATION

Pongtep Angkititrakul and John H.L. Hansen

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

ABSTRACT

In this paper we study the problem of identifying in-set and out-of-set speakers. The goal is to identify whether an unknown input speaker belongs to either a group of in-set speaker or an unseen out-of-set group. A state-of-the-art GMM classifier with Universal Background Model (UBM), and standard likelihood ratio test are used as our baseline system. We propose an alternative hypothesis testing method that employs neighborhood information with respect to each in-set speaker model in the model space based on the Kullback-Leibler divergence. The Bayes Factor is used in the verification stage (accept/reject hypothesis). We evaluate the proposed procedure on a clean CORPUS1 set, and a noisy CORPUS2 set which contains session-to-session variability. Experiments show an improvement in Equal Error Rate for the system even when in-set speaker models are acoustically close in the model space, and as the in-set speaker size increases.

1. INTRODUCTION

The problem of identifying in-set/out-of-set speakers (or open-set speaker recognition) is a major challenge when the subjects grouped for the in-set are arbitrary. If the in-set group possesses some physical trait (i.e., age, physical size, etc.) or language trait (i.e., same geographical region, dialect, accent, etc.), the in-set speaker detection is presumably easier. However, if testing utterances are unconstrained, a testing observation may or may not be one of the defined classes in the training data set. In general, based on the likelihood of the speech features, open-set speaker recognition first classifies the observation into the most likely speaker class from a set of *known* speakers, as the "closed-set" speaker identification. Nevertheless, open-set speaker recognition has to further make a decision to either accept or reject a potential speaker, whether the observation really belongs to one of the in-set (enrolled/target) speakers of the group, or out-of-set speakers (impostors/garbage/outliers). Typical applications of open-set speaker recognition includes multi-user information access devices, forensic speaker identification, and others.

Some prior studies of open-set speaker recognition examined new decision procedures, which were based on the score space as the outcome of the classifiers (i.e., GMM-based model [3] or VQ-based model [2]). Various score normalization techniques have been compared in [8].

In this paper, we address an alternative hypothesis testing based on Bayes Factors (BF) and neighborhood information in the model space [5, 4] for open-set recognition problem. The experiments are constructed for a text-independent task with limited training data for each speaker model (approximately five seconds worth of speech for each speaker). A fixed-size speaker set was randomly selected as the working speaker space, when each speaker rotated its role acting as in-set or out-of-set speakers for different experiments. We use GMM/UBM based classifier as our baseline system. The distance measure between neighbors is based on the relative information entropy, Kullback-Leibler (KL) divergence. An alternative Bayesian approach is also compared with the conventional Likelihood Ratio Test (LRT). We believe there will be a reduction in the Equal Error Rate (EER) of the system performance with the new Bayesian approach compared with the standard approach based on UBM, when the speaker models are not well separated in the model space.

This paper is organized as follows: First, we discuss the objective formulation of the open-set problem. In Sec.3, we briefly review the GMM-UBM based classifier. Next, we introduced the speaker's neighborhood information within the model space in Sec.4. In Sec.5, we discuss Bayesian Interpretation and alternative hypothesis testing. We explain and report our experimental results in Sec.6. Finally, conclusions and future work are discussed in Sec.7.

2. OBJECTIVE FORMULATION

We assume we are given a set of N in-set (enrolled) speakers in a system, and the collected data $\mathbf{X}_n$, corresponding to each enrolled speaker S_n, $1 \leq n \leq N$. Let the data $\mathbf{X}_0$ represent all other non-enrolled speakers in the development set. Each speaker dependent statistical model $\{\Lambda_n \in \mathbf{\Lambda}, 1 \leq n \leq N\}$ can be obtained from $\{X_{n1}, X_{n2}, \ldots, X_{nT_n}\}$ where T_n denotes the total number of samples belong to speaker S_n.

If $\mathbf{X}$ denotes the sequence of feature vectors extracted from the test utterance, then the problem of identifying in-set versus out-of-set speaker requires that we perform two statistical stages. In the first stage, called *speaker identification* or *speaker classification*, we first classify $\mathbf{X}$ into one of the most likely in-set speakers, Λ^*, e.g.,

$$\Lambda^* = \underset{1 \leq n \leq N}{argmax}\, p(\mathbf{X}|\Lambda_n). \quad (1)$$

In the second stage, called *speaker verification* or *outlier verification*, we verify whether the observation $\mathbf{X}$ truly belongs to Λ^* or not (accept/reject). In general, this stage is formulated as a problem of statistical hypothesis testing when the *null* hypothesis $\mathbf{H_0}$, represents the hypothesis that $\mathbf{X}$ really belongs to model Λ^*, against the competitive hypothesis $\mathbf{H_1}$, that represents the hypothesis where $\mathbf{X}$ is actually "not" from model Λ^*. The likelihood

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

ratio test is given by:

$$\frac{p(\mathbf{X}|\Lambda^*)}{p(\mathbf{X}|\Lambda_0)} \begin{cases} \geq \gamma & : \text{accept } H_0, \\ < \gamma & : \text{reject } H_0 \text{ (accept } H_1). \end{cases} \quad (2)$$

where γ is a threshold, Λ_0 is a competitive model (*anti-model*), and $p(\cdot|\cdot)$ is the likelihood generated from each model. In practice, it is impossible to have a true anti-model for the competitive speaker class, otherwise we could define such a speaker model as one class in the training phrase. The conventional strategy assumes another special class, or speaker independent model, as a universal representative of all other speakers excludes all the in-set speakers (e.g., UBM).

3. GMM-UBM AND MAP ADAPTATION

Recently, state-of-the-art Gaussian Mixture Model (GMM) with Maximum A Priori (MAP) speaker adaptation has become the dominant approach in text-independent speaker recognition [7]. A speaker independent model, or Universal Background Model (UBM), is trained from the non-target speakers by the Expectation Maximization (EM) algorithm. The probability density function (*pdf*) of an M-Gaussian components for D-dimensional observation vectors $\mathbf{X}$ is defined as:

$$p(\mathbf{X}|\Lambda_0) = \sum_{m=1}^{M} \omega_{0m} G_{0m}(\mathbf{X}), \quad (3)$$

where ω_{0m} is the weight of the m-th component, and G_{0m} is the Gaussian probability density function with mean μ_{0m} and covariance matrix Σ_{0m}, which is usually assumed diagonal. For each target speaker, a speaker dependent GMM ($\Lambda_n : \{\hat{\omega}_n, \hat{\mu}_n, \hat{\Sigma}_n\}$) can be created by MAP adaptation of UBM parameters $\{\omega_0, \mu_0, \Sigma_0\}$ and the training data $\mathbf{X}_n$ via the following formula:

$$\hat{\mu}_{nm} = \frac{\eta_m}{\eta_m + r} E_m(\mathbf{X}_n) + \frac{r}{\eta_m + r} \mu_{0m}, \quad (4)$$

where η_m is the weight assigned to the m-th component in the UBM, and r is a relevance factor which depends on the parameter and controls the balance of adaptation. The speaker dependent model obtained from MAP-adapted UBM provides a tighter coupling between the speaker specific model and the UBM. For our study, only mean adaptation is considered since out prior experiments showed this was superior to the system with all parameter adapted.

4. NEIGHBORHOOD IN MODEL SPACE

The basic idea behind "nested-neighborhood" structure in the model space is that all competing models of a given model sit inside one neighborhood of the underlying model. The small neighborhood can be viewed as a robust representation of the original model and it contains all possible variants from the original model due to mismatches and other estimation errors, when the large neighborhood represents all potential competing models of the original model. The idea has been successfully applied in the HMM model space for Utterance Verification (UV) applications [5]. We developed such an idea for open-set recognition in this section.

Given a set of N in-set speaker GMMs in the system, denoted as $\{\Lambda_n, 1 \leq n \leq N\}$, let each Λ_n can be viewed as a point in the model space Λ. For any given two speaker models in the space and their corresponding training data, we can estimate a distance between these two GMMs using the symmetric Kullback-Leibler (KL) divergence. The symmetric KL divergence is defined as the sum of relative entropy between model Λ_i and model Λ_j, and also between model Λ_j and model Λ_i [1]:

$$KL(\Lambda_i, \Lambda_j) = E_{\Lambda_i(\mathbf{X})}[log\frac{\Lambda_i(\mathbf{X})}{\Lambda_j(\mathbf{X})}] + E_{\Lambda_j(\mathbf{X})}[log\frac{\Lambda_j(\mathbf{X})}{\Lambda_i(\mathbf{X})}] \quad (5)$$

where $\Lambda_i(\mathbf{X})$ and $\Lambda_j(\mathbf{X})$ are the likelihoods of occurances of observation $\mathbf{X}$, given that it belongs to model Λ_i and Λ_j respectively. The KL divergence quantifies the information for discriminating between the two speaker models. Subsequently, we can construct an $N \times N$ distance matrix where diagonal elements are zeros, denoted as Δ. The distance matrix is then normalized by its maximum element. Next, for each speaker model Λ_n, we label all in-set speaker models for a nested neighborhood (NNB) with respect to model Λ_n, with the following definition [5]:

- *Self neighborhood* $\Lambda_n^{(0)}$: consists of only the model Λ_n itself.
- *1st-level neighborhood* $\Lambda_n^{(1)}$: is a small neighborhood that surrounds the model Λ_n. The neighborhood consists of speaker models Λ_i such that $0 < \Delta[n,i] \leq \alpha$, for $1 \leq i \leq N$.
- *2nd-level neighborhood* $\Lambda_n^{(2)}$: is a medium neighborhood that surrounds the 1st-level neighborhood. The neighborhood consists of speaker models Λ_i such that $\alpha < \Delta[n,i] \leq \beta$, when $\beta > \alpha$, for $1 \leq i \leq N$.
- *3rd-level neighborhood* $\Lambda_n^{(3)}$: is a larger neighborhood surrounding the 2nd-level neighborhood. The neighborhood consists of speaker models Λ_i such that $\Delta[n,i] > \beta$, for $1 \leq i \leq N$.

Here, we have α and β as distance bounds, and are constrained as $0 < \alpha < \beta < 1.0$. Both parameters can be set dependent on the in-set speakers model space.

With the use of neighborhood information, we can define an alternative hypothesis testing as:
$\tilde{H}_0$: The observation belongs to the self neighborhood or 1st-level neighborhood
$\tilde{H}_1$: The observation belongs to the 2nd-level neighborhood

5. BAYESIAN INTERPRETATION

In the Bayesian framework, we first estimate an *a priori* pdf for each class. Next, the decision is made based on the calculation of the *Bayes Factors* [5, 6]. Given the observation vectors $\mathbf{X}$ along with two hypotheses H_0 and H_1, the Bayes Factor is computed as:

$$BF = \frac{\hat{p}(\mathbf{X}|H_0)}{\hat{p}(\mathbf{X}|H_1)} = \frac{\int p(\mathbf{X}|\Lambda_0, H_0) \cdot p(\Lambda_0|H_0) d\Lambda_0}{\int p(\mathbf{X}|\Lambda_1, H_1) \cdot p(\Lambda_1|H_1) d\Lambda_1} \quad (6)$$

where Λ_0, Λ_1 are the model parameters under H_0 and H_1 respectively, $p(\Lambda_0|H_0)$ and $p(\Lambda_1|H_1)$ are the prior densities, and $p(\mathbf{X}|\Lambda_0, H_0)$ and $p(\mathbf{X}|\Lambda_1, H_1)$ are the likelihood functions of the model parameters under their hypotheses.

Bayes Factor is the ratio of the posterior odds [1] of H_0 to its prior odds, regardless of the value of the prior odds [6]. Therefore,

[1] Any probability can be converted to the odds scale as $odds = probability/(1 - probability)$.

Bayes Factors can be used to compare with a threshold to make a decision with regards to H_0 [5]. Based on neighborhood definition, The Bayes Factors used to verify the hypotheses $\tilde{H}_0$ and $\tilde{H}_1$ can be simplified as:

$$\xi = \frac{a \cdot \sum_{\Lambda_i \in \Lambda_n^{(1)}} p(\mathbf{X}|\Lambda_i) \cdot p(\Lambda_i|\Lambda_n^{(1)}) + b \cdot p(\mathbf{X}|\Lambda_n^{(0)})}{\sum_{\Lambda_i \in \Lambda_n^{(2)}} p(\mathbf{X}|\Lambda_i) \cdot p(\Lambda_i|\Lambda_n^{(2)})} \quad (7)$$

where ξ is a threshold, $p(\mathbf{X}|\Lambda_i)$ is the likelihood of observation $\mathbf{X}$ given that it belongs to model Λ_i, $p(\Lambda_i|\Lambda_n^{(1)})$ and $p(\Lambda_i|\Lambda_n^{(2)})$ are the prior probabilities with constraints that $\sum_{\Lambda_i \in \Lambda_n^{(1)}} p(\Lambda_i|\Lambda_n^{(1)}) = 1$ and $\sum_{\Lambda_i \in \Lambda_n^{(2)}} p(\Lambda_i|\Lambda_n^{(2)}) = 1$, and a, b are tuning weights.

6. EXPERIMENTS

6.1. Experimental Setup

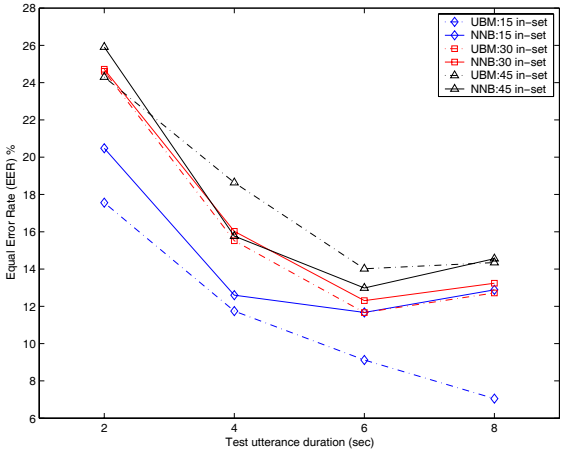
The two corpora are used for our study. While CORPUS1 has only a single recording session and is noise-free, CORPUS2 has multiple recording sessions and is noisy.

6.1.1. CORPUS1: Clean speech corpus (TIMIT)

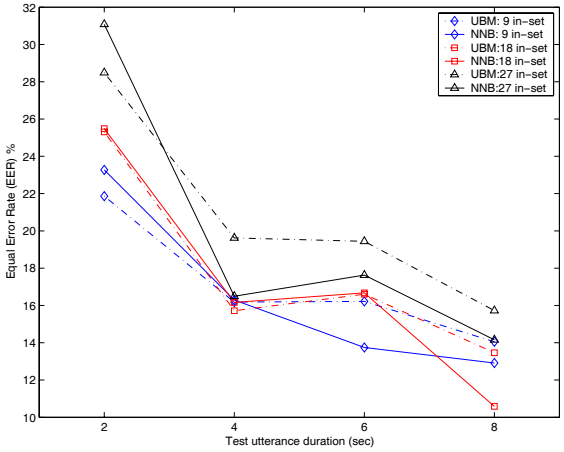
A set of 60 male speakers was randomly selected as a speaker sample space. These 60 speakers serve both as in-set speakers and out-of-set speakers (impostors) depending on the experimental set. In particular, three different sizes of in-set speakers are considered (e.g., 15, 30, and 45). For example, 15 speakers were randomly selected from the speaker sample space as the in-set speakers, with the remaining 45 speakers taking the role of impostors ('15in/45out'). Similar to other Round-Robin test procedures, different combinations of in-set and out-of-set speakers were also selected, resulting in four distinct '15in/45out' groups, two distinct '30in/30out' groups, and two (with some overlap) '45in/15out' groups. The training and testing speech data of each speaker were randomly selected and concatenated from the original TIMIT database, with no data overlap. The training data was limited to approximately 5 seconds worth of speech, while testing data was created for 2, 4, 6, and 8 seconds worth of speech. For more reliable results, similar to the above setup, another training and testing data set were also created for the same 60 speakers for comparison. Also, some speakers excluded from the working speaker sample space are used as development data.

6.1.2. CORPUS2: Noisy speech corpus

The second data test set consists of speech recorded with aircraft cockpit noise. The transmissions are short in duration and have multiple recording phases (i.e., contains session-to-session variability). Similar to the experimental framework for the CORPUS1, a collection of 36 speakers were randomly selected as the speaker sample space. The size of different in-set/out-of-set speaker groups are '9in/27out'(4 groups), '18in/18out'(2 groups), and '27in/9out' (2 groups). The training data was also limited to approximately 5 seconds worth of speech, and testing data was again created for 2, 4, 6, and 8 seconds worth of speech. Two sets of training and testing for each speaker group were also created for overall averaging of the results. Finally, we tested a portion of the noisy CORPUS2 data using NIST STNR tool, and found that our data has a STNR value of 19 dB, which is much more noisy than CORPUS1, which has a value of 39 dB.



(a) CORPUS1: Clean speech corpus results.



(b) CORPUS2: Noisy speech corpus results.

Fig. 1. *In-set vs Out-of-set Speaker Identification performance in terms of EER(%) at 2, 4, 6, and 8 second test utterances.*

6.2. Front-end Processing

The speech analysis frame rate is set to 30 ms with a 10 ms skip rate. Speech is pre-emphasized with the filter $(1 - 0.95z^{-1})$. Nineteen-dimensional Mel-Frequency Cepstral Coefficients (MFCC) are extracted, and appended with delta-energy. For CORPUS1, silence and low-energy speech parts are removed using a general energy detection technique (e.g., frames that have higher energy than the pre-defined threshold are selected). For CORPUS2, frame selection is based on formant information (e.g., frames that have the estimated three formant locations lie within a specified frequency range (200-3900 Hz) are selected). Cepstral Mean Normalization is applied to each utterance to reduce channel based spectral shaping.

6.3. Evaluations

6.3.1. Baseline System

First, the UBM is constructed from speakers in the development set, with 32 Gaussian components. The GMM construction starts with vector quantization codebooks with several updated itera-

tions, and the GMM parameters are consequently adjusted with EM iterations. A single speaker-dependent GMM for each in-set speaker is then estimated from the UBM, based on MAP adaptation. The number of Gaussian mixtures is also fixed to 32 for all speakers in our experiments. The baseline system employs equation (1) and (2), when the verification stage is decided based on likelihood ratio test against the UBM ($\Lambda_0 = \Lambda_{UBM}$).

6.3.2. Neighborhood Information and Bayesian Approach

We keep the same set of speaker GMMs as obtained from the baseline system. During the training stage, the distance matrix of each in-set speaker group is computed based on the KL divergence(e.g., from Eq. 5). Each speaker now has a codebook that contains speaker neighborhood information, along a with speaker GMM. For each testing observation, the most likely in-set speaker is again chosen from the maximum likelihood score. The verification stage employs the alternative hypothesis as mentioned in Eq. 7.

6.3.3. Results and Discussions

Figure 1 shows the average Equal Error Rate (EER) of the system performance over all experiments with the same 'in-set/out-of-set' size at different testing utterance durations. 'UBM' (dash lines) denotes the baseline system, which tests the likelihood ratio against the UBM. 'NNB' (solid lines) denotes the system with alternative hypothesis testing based on nested neighborhood information. As we can see, the EERs tend to degrade as the size of the in-set speaker group increases. For the CORPUS2 results, the new hypothesis testing performs slightly worse for the 2 second test utterance, but dramatically reduces EERs as the duration increases (e.g., 4 seconds and greater). The NNB also enhances the performance of larger sized in-set groups with performance approaching that seen for smaller groups. For the clean CORPUS1, we do not see consistent improvement when employing our alternative hypothesis testing (NNB), except some EER reduction for experiments with in-set size 45 for 4 and 6 second test utterances.

It would be useful to consider performance differences using the NNB hypothesis test for the clean CORPUS1 and noisy CORPUS2 experiments. The reason is that for the clean CORPUS1 set, the speaker models are well-trained and more separated than speaker models seen in the CORPUS2 set. Figure 2 shows an example histogram of the distance distribution of one experiment with '15in/15out' clean CORPUS1 and '18in/18out' noisy CORPUS2. For CORPUS2, the distance distribution is well distributed on the '1st-level neighborhood' and '2nd-level neighborhood', probably because of either (i) the physical properties of selected speakers; or (ii) the less discriminative models as a result of noise content in the data. Such a distribution can cause more confusion for the "closed-set" recognition, but is useful for NNB hypothesis testing. For CORPUS1, the distance distribution is more distributed on the '2nd-level neighborhood' (i.e., 0.2-0.55 for this example) and '3rd-level neighborhood' (i.e., 0.55-1.0), which shows good discriminative ability between speaker models. So, for this in-set group, few or none of the speakers are in the 1st-level neighborhood, and therefore the NNB-nested neighborhood method will not be as successful. Under this condition, the NNB approach becomes similar to cohort normalization (CN) [8] when the cohort speakers are selected from the in-set speaker group. This observation also helps to explain why improvement occurs for the CORPUS1 '45in/15out' test configuration, since a larger number of in-set speakers fall within the 1st-level neighborhood.

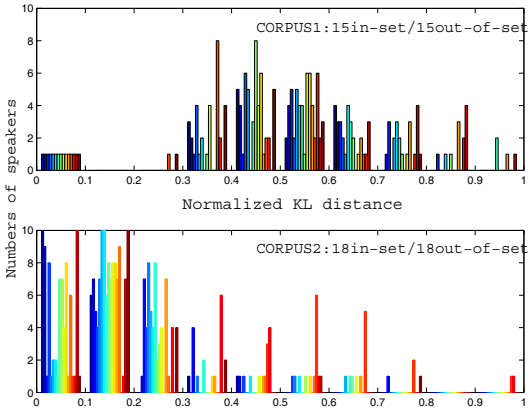


Fig. 2. *Histogram of the distance distributions: (a) CORPUS1 contains clean data, (b) CORPUS2 contains noisy data with session-to-session variability.*

Therefore, a distance based analysis of the in-set speaker models as shown in the distributions from Fig. 2 can be an effective way of predicting when NNB versus traditional hypothesis testing will be more successful.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have studied the problem of identifying in-set versus out-of-set speakers. A state-of-the-art GMM-UBM system with MAP adaptation, and standard likelihood ratio test was used as our baseline system. We proposed the NNB (nested neighborhood) method which employs neighborhood information in the model space and the Bayes Factors. Such hypothesis testing is promising when the speaker size of the in-set group increases, or speaker models are close together in the model space and distance distribution (e.g., Fig. 2) is well distributed. The NNB method was evaluated on both clean (CORPUS1) and noisy (CORPUS2) data. Our future work will consider a feature fusion technique with the use of confidence measures to improve the system performance.

8. REFERENCES

[1] Mathieu Ben and Frédéric Bimbot, "D-MAP: A distance-normalized MAP estimation of speaker models for automatic speaker verification," in *Proc. IEEE ICASSP 2003*, Hong Kong, April 2003, pp. II.69-72.

[2] Jiuqing Deng and Qixiu Hu, "Open set text-independent speaker recognition based on set-score pattern classification," in *Proc. IEEE ICASSP 2003*, Hong Kong, April 2003, pp. II.73-76.

[3] Yifan Gong, "Noise-robust open-set speaker recognition using noise-dependent Gaussian mixture classifier," in *Proc. IEEE ICASSP 2002*, Orlando, Florida, May 2002, pp. I.133-136.

[4] Hui Jiang and Li Deng, "A Bayesian Approach to the Verification Problem: Applications to Speaker Verification," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 8, pp. 874-884, Nov 2001.

[5] Hui Jiang and Chin-Hui Lee, "A New Approach to Utterance Verification Based on Neighborhood Information in Model Space," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 5, pp. 425-434, Sept 2003.

[6] Robert E. Kass and Adrian E. Raftery, "Bayes Factors," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 773-795-434, June 1995.

[7] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing.*, vol. 10, no. 1-3, pp. 19-41, Jan./Apr./July 2000.

[8] P. Sivakumaran, J. Fortuna, and A.M. Ariyaeeinia, "Score Normalization Applied to Open-Set, Text-Independent Speaker Identification," in *EUROSPEECH/INTERSPEECH 2003*, Geneva, Switzerland, Sept 2003, pp. 2669-2672.