# Phonetic Speaker Recognition

*Walter D. Andrews, Mary A. Kohler, and Joseph P. Campbell*

Department of Defense
Speech Processing Research
waltandrews@ieee.org, m.a.kohler@ieee.org, j.campbell@ieee.org

## Abstract

This paper introduces a novel language-independent speaker-recognition system based on differences among speakers in dynamic realization of phonetic features (i.e., pronunciation) rather than spectral differences in voice quality. The system exploits phonetic information from six languages to perform text independent speaker recognition. All experiments were performed on the NIST 2001 Speaker Recognition Evaluation Extended Data Task. Recognition results are provided for each of the six language front ends and for various fusions. The fusion results demonstrate that speaker recognition capability for speech in languages outside the system is successful.

## 1 Introduction

Most practical methods of speaker recognition, and especially those with very limited training, are based on differences in voice quality, broadly defined, rather than on the phonetics of pronunciation [1], [2]. Although there can be little doubt that the dynamics of pronunciation contribute to human recognition of speakers, exploiting such information automatically is difficult because, in principle, comparisons must be made between different speakers saying essentially "the same things."

One technique to do this would be to use speech recognition to capture the exact sequence of phones, examine the acoustic phonetic details of different speakers producing the same sounds and sequences of sounds, and compare these details across speakers or score them for each speaker against a model.

As an extreme example, given speakers *A*, *B*, and *C*, where speaker *A* lisps and speaker *B* stutters; then given perfect recognition of a large enough sample of speech by all three, the acoustic scores of the [s] and [sh] sounds might distinguish *A* from *B* and *C*, and either the acoustic scores or the HMM path traversed by the initial stop consonants, for example, might distinguish *B* from *C* and *A*.

An obvious problem with this approach is that recognizers are usually optimized for recognition not of phones but of words, use powerful word n-gram statistics to guide their decisions, and train their acoustic processing, model topologies, and time alignment to ignore speaker differences.

What we need is a tool, which will consistently recognize and classify as many phonetic states as possible, regardless of their linguistic roles (i.e., what words are being spoken), using sufficiently sensitive acoustic measurements, so that comparisons can be made among different speakers' realizations of the "same" speech gestures.

We develop a speaker-recognition system based only on phonetic sequences instead of the traditional acoustic feature vectors. Although the phones are generated based on the acoustic feature vectors, the recognition is performed strictly from the phonetic sequence created by the phone recognizer(s). Speaker recognition is performed using six phone recognizers, which were trained on six languages. Recognition of the same speech sample by the six recognizers constitutes six different "views" of the phonetic states and state sequences uttered by the speakers.

These six independent systems are fused using a simple linear combination. We show that fusing the systems in this manner provides a significant decrease in equal error rate (EER). We also show that there is no significant loss in performance if the language of the speaker in question is not directly modeled by the system.

## 2 NIST Extended Data Task

All of the experiments in this paper use the data from the NIST 2001 Speaker Recognition Evaluation Extended Data Task. NIST's purpose in creating this task was to promote the exploration and development of new approaches to the speaker recognition challenge, such as the idiolectal characteristics reported in [3].

In previous evaluations, the one speaker detection task was viewed as a limited training data task, i.e., only two minutes of training data were provided for each of the hypothesized speakers. For the 2001 evaluation, the entire Switchboard-I corpus was prepared for the Extended Data Task. Along with the audio data, NIST provided both automatic speech recognition transcriptions, courtesy of Dragon Systems, and manual transcripts for the entire corpus. All forms of data were permitted for training speaker models either alone or in combination.

The speaker model training data was comprised of one, two, four, eight, and sixteen conversations. NIST employed a jackknife approach to rotate through the training and testing conversations to insure there is an adequate number of tests. Table I provides a breakdown, based on the number of training conversations, of the NIST Extended Data Task.

Table I: NIST Extended Data Task

| Number of Training Conversations | Number of Unique Speakers | Number of Test Conversations |
|---|---|---|
| 1 | 483 | 16429 |
| 2 | 442 | 15363 |
| 4 | 385 | 13777 |
| 8 | 273 | 10377 |
| 16 | 57 | 2696 |
| Total | 483 | 58642 |

For testing, the same options were available as in training. The recognition feature could be computed from either the acoustic data, the transcriptions or a combination

of both. The number of test conversations for each set of training conversations is provided in Table I. The test set contains matched handset and mismatched handset conditions as well as a few cross-gender trials.

## 3 Phonetic Speaker Recognition

Phonetic speaker recognition is performed in four steps. First, a phone recognizer, in the appropriate language, processes the test speech utterance to produce phone sequences. Then a test speaker model is generated using phone n-gram (n-phone) frequency counts. Next, the test speaker model is compared to the hypothesized speaker models and the Universal Background Phone Model (UBPM). Finally, the scores from the hypothesized speaker models and the UBPM are combined to form a single recognition score.

The single-language system is generalized to accommodate multiple-languages by incorporating phone recognizers trained on several languages resulting in a matrix of hypothesized speaker models. The system here used *P* phone recognizers and *P* UBPMs, one UBPM for each phone recognizer. (The use of a single integrated UBPM will be reported on later.) With *M* hypothesized speakers, the multilanguage system produces *PxM* hypothesized speaker models and scores. Figure 1 shows this multilanguage phonetic speaker-recognition system. The following sections provide more details for the modeling and recognition process.

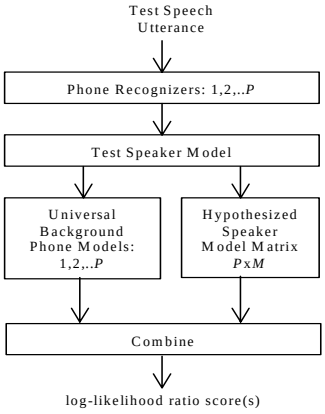


*Figure 1. Multilanguage Phonetic Speaker-Recognition system*

### 3.1 Phone Recognition

The phone recognition process takes advantage of the algorithm that Zissman created for Parallel Phone Recognition with Language Modeling (PPRLM) [4], since this recognizer was created solely for phone recognition. This algorithm calculates twelve cepstral $(c_1 - c_{12})$ and thirteen delta-cepstral $(c'_0 - c'_{12})$ features on 20 ms frames with 10 ms updates. The cepstra and delta-cepstra are sent as two independent streams to fully connected, three-state, null-grammar HMMs.

The HMMs were trained on phonetically marked speech from the OGI multilanguage corpus in six languages: English (EG), German (GE), Hindi (HI), Japanese (JA), Mandarin (MA), and Spanish (SP). The corpus was hand-marked by native speakers in each language using OGI symbols for two of the languages and Worldbet symbols for the remainder. The number of phonetic symbols differs for each language from 27 for Japanese to 51 for Hindi, and includes one symbol to represent silence.

The algorithm uses a Viterbi HMM decoder implemented with a modified version of the HMM Toolkit. The output probability densities for each observation stream (cepstra and delta-cepstra) in each state are modeled as six univariate Gaussian densities. The output from the HMM recognizer for each language provides four estimates: the symbol for the recognized phone, its start time, its stop time, and its log-likelihood score. For this paper we only used the recognized phone although future plans include exploiting the other estimates.

There are a number of variations for formatting the output phones from the recognizer. Doddington showed for word n-grams that including start and stop tags improved speaker recognition performance [3]. In this paper, we naively added start and stop tags based on pairs of silence phone labels. All phones between two silence phone labels were considered an utterance.

### 3.2 Hypothesized Speaker Model

As noted in section 2, a jackknife scheme determined the amount of training and testing data for the extended training task. NIST provided a control file listing hypothesized and test speakers, along with a training and testing conversation list [5]. The list provided training information for one, two, four, eight, and sixteen conversations. As a result, a particular hypothesized speaker will have multiple models for a given test set.

Speaker dependent language models, *H*, are generated using a simple n-phone frequency count for each language and consist of all the unique n-phones with the corresponding frequency counts for a given speaker. Unlike the state-of-the-art GMM-UBM systems, the speaker models are not adapted from the UBPM.

### 3.3 Universal Background Phone Model

The UBPM, *U*, is generated as determined by the NIST control file (specified in [5]), which provides a list of hypothesized and test speakers for exclusion from the UBPM. All of the conversations for the remaining speakers were used to build the UBPM using n-phone frequency counts. For this paper, each of the six phoneme recognizers has a corresponding independent UBPM.

### 3.4 Test Speaker Model

A test set is specified in the NIST control file for all hypothesized speaker models. The test set contains true speaker trials, impostor trials, matched handset, mismatched handset, and a few cross-gender trials. Once the test speech utterance to be tested is processed by the phone recognizer(s), a test speaker model, *T*, is generated using n-phone frequency counts. Doddington improved performance by ignoring infrequent word n-grams [3]. This is also the case with the phonetic approach. In this paper we used triphone $(n = 3)$ models and ignore n-phones that occur less than 1,000 times, which we refer to as $c_{\min}$.

### 3.5 Combining Scores

For a single-language phonetic speaker-recognition system, the scores from the hypothesized speaker models and the UBPM are combined to form the recognition score $\eta_i$ using a generalized conventional log-likelihood ratio given by

$$\eta_i = \frac{\sum_n \left( w(n) \left[ S_i(n) - B(n) \right] \right)}{\sum_n w(n)}$$

where $n$ is a n-phone type corresponding to the test speaker model, $T$, and the sums run over all of the n-phone types in the test speaker model, $T$. $S_i$ represents the log-likelihood score from the $i^{th}$ hypothesized speaker model, $H_i$, and $B$ is the log-likelihood score from the UBPM, $U$, for the n-phone type, $n$. The log-likelihood scores $S_i$ and $B$ are defined by

$$S_i(n) = \log\left[ \frac{H_i(n)}{N_{H_i}} \right] \quad and \quad B(n) = \log\left[ \frac{U(n)}{N_U} \right],$$

where $N_{H_i}$ and $N_u$ represent the total number of unique n-phone types in the $i^{th}$ hypothesized speaker model and UBPM, respectively. $H_i(n)$ and $U(n)$ represent the number of occurrences of a particular n-phone type, $n$, in the hypothesized speaker model and UBPM, respectively.

The weighting function $w(n)$ is based on the n-phone token count, $c(n)$, and the discounting factor, $d$. The n-phone token count, $c(n)$, corresponds to the number of occurrences of a particular n-phone type $n$ in the test speaker model, $T$. The weighting function, which could be made language dependent, is given by

$$w(n) = c(n)^{1-d} .$$

The discounting factor, $d$, has permissible values between 0 and 1. When $d = 1$ a complete discounting occurs, resulting in $w(n) = 1$. This gives all n-phone types the same weight regardless of the number of occurrences in the test speaker model, $T$. When $d = 0$, all n-phone types are weighted by their number of occurrences in the test speaker model, $T$.

The scores from each of the single-language phonetic speaker-recognition systems can be fused by a simple linear combination

$$\lambda_i = \sum_j^P \alpha_j \eta_{j,i} .$$

where $\alpha$ are the language dependent weights.

## 4 Results

### 4.1 Individual Languages

Figure 2 provides the detection-error trade-off (DET) curves for each of the language-dependent phonetic speaker-recognition systems. The curves present results for speakers trained with sixteen conversations, triphone models, $d = 1$, and $c_{\min} = 1{,}000$ (only triphone types occurring 1,000 times or more are considered). For the single-language recognition system, only the selected language recognizer's weight, $\alpha$, is nonzero. As one might expect, the phonetic speaker-recognition system using the English phone recognizer performed best. It is interesting to note that processing English utterances with non-English phone recognizers does not significantly degrade performance of the speaker recognition.

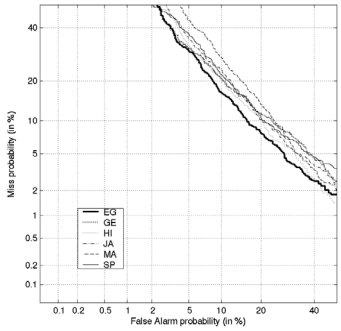


*Figure 2. DET Curves for each language*

### 4.2 Fusion

The first step toward producing a language-independent speaker-recognition system is the fusion of the six language-dependent phonetic speaker-recognition systems. The experiment was performed with triphone models, $d = 1$, and $c_{\min} = 1{,}000$; with the language dependent weights $\alpha = 1/6$, thus, giving each language-dependent score equal weighting. Figure 3 shows the result of this system, grouped into the number of conversations used for training. The results clearly indicate there is significant improvement with the fused system over the single-language systems when sixteen conversations are used for training.

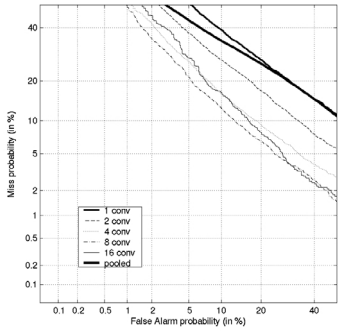


*Figure 3. DET Curve for all six languages combined using equal weighting*

The next step is to test the language independence of the multilanguage phonetic speaker-recognition system on the NIST Extended Data Task, which contains only English. This experiment gave zero weight to the English phone recognizer. The remaining five recognizers were fused with

equal weight. Figure 4 shows the DET curves for this system. The system used: triphone models, $d = 1$, and $c_{\min} = 1,000$. Removing the English phone system results in only a slight degradation in performance (for training with sixteen conversations), thus, supporting our claim of a language-independent phonetic speaker-recognition system (at least for English input speech).

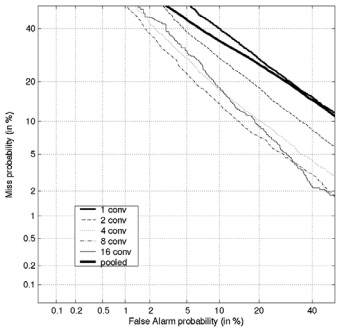


*Figure 4. DET Curve for system without EG phone recognizer*

### 4.3 Speaker Entropy

One method for determining the power of phonetic-based speaker recognition is to analyze the speaker entropy. Figure 5 shows a speaker entropy scatter plot for the triphones from the EG phone recognizer. The speaker entropy is computed as in [3] by

$$H_n = \sum_m^M \left( -\sum_n \left( P_m(n) \log\left[ P_m(n) \right] \right) \right).$$

*P* is the ratio of the number of occurrences of a particular n-phone type, *n*, for a given speaker, *m*, and the total number of occurrences of the particular n-phone type, *n*, for all *M* hypothesized speakers. The speaker entropy is plotted against the frequency count of triphones in the NIST Extended Data Task.

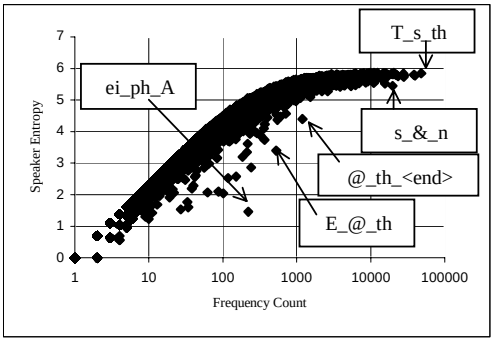


*Figure 5. Speaker entropy plot for triphones*

This result is similar to that of [3], which is based on word n-grams. We are interested in n-phones that have a high occurrence and a low speaker entropy value, so the most interesting points on the speaker entropy plot are the outliers.

## 5 Conclusions

We have shown that there is indeed speaker recognition power at the phonetic level. Six language-dependent phonetic speaker-recognition systems were developed using English, Spanish, Hindi, Japanese, Mandarin, and German training. We showed that individually all of these systems performed similarly on the NIST Extended Data Task, which has only English speech.

Next we developed a system that combined all six of the language dependent systems using a language-dependent weighting function. The fusion resulted in an increase in performance over that of the six individual systems.

We then showed that removing the system containing the language of interest results in only a slight performance degradation, thus proving the concept of a language-independent phonetic speaker-recognition system (for English).

The concept of language-independent phonetic speaker-recognition is in its infancy. It ushers in an entirely new approach to speaker recognition that has recently been dominated by the GMM-UBM. There are a number of areas that require further research, such as reduced training requirements, improved n-phone models, improving phone estimation, more sophisticated techniques for fusing, duration tagging of phones, gender dependent phone models, integrated UBPM, using tokens more general than phones, and fusion with word-based and/or GMM-UBM systems [6].

## 6 Acknowledgements

The authors thank George Doddington and John Godfrey for their helpful discussions.

## 7 References

[1] Reynolds, D., T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41.

[2] Weber, F., B. Peskin, et al., "Speaker Recognition on Single- and MultiSpeaker Data," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 75-92.

[3] Doddington, G., "*Some Experiments on Idiolectal Differences Among Speakers*," http://www.nist.gov/speech/tests/spk/2001/doc/, January 2001.

[4] Zissman, M., "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. On SAP*, vol. 4, Issue 1, January 1996.

[5] Przybocki, M., and A. Martin, "*The NIST Year 2001 Speaker Recognition Evaluation Plan*," http://www.nist.gov/speech/tests/spk/2001/doc/, March 1, 2001.

[6] Andrews, W., M. Kohler, and J. Campbell, "Acoustic, Idiolectal, and Phonetic Speaker Recognition," To appear, *Proceedings of 2001: A Speaker Odyssey, The Speaker Recognition Workshop*, Chania, Crete, Greece, June 18-22, 2001.