CONDITIONAL PRONUNCIATION MODELING IN SPEAKER DETECTION

David Klusáček
Charles University
Department of Mathematics and Physics
klusacek@atrey.karlin.mff.cuni.cz

Jiří Navrátil
IBM
T.J.Watson Research Center
jiri@us.ibm.com

D. A. Reynolds, J. P. Campbell
MIT
Lincoln Laboratory
{dar,jpc}@ll.mit.edu

ABSTRACT

In this paper, we present a conditional pronunciation modeling method for the speaker detection task that does not rely on acoustic vectors. Aiming at exploiting higher-level information carried by the speech signal, it uses time-aligned streams of phones and phonemes to model a speaker's specific pronunciation. Our system uses phonemes drawn from a lexicon of pronunciations of words recognized by an automatic speech recognition system to generate the phoneme stream and an open-loop phone recognizer to generate a phone stream. The phoneme and phone streams are aligned at the frame level and conditional probabilities of a phone, given a phoneme, are estimated using co-occurrence counts. A likelihood detector is then applied to these probabilities.

Performance is measured using the NIST Extended Data paradigm and the Switchboard-I corpus. Using 8 training conversations for enrollment, a 2.1% equal error rate was achieved. Extensions and alternatives, as well as fusion experiments, are presented and discussed.

1. INTRODUCTION

For automatic speaker recognition systems, it is not *what* you say but *how* you say it that is important. The particular content being conveyed is not as important as how the words sound (i.e., pronunciations) or even the selection and combination of words used (i.e., idiolect). In most automatic speaker recognition systems, however, only the general underlying sounds of a person's voice are modeled and represented via Gaussian Mixture Models (GMM) and short-term acoustic features [4], providing only indirect and implicit modeling of pronunciations. For text-dependent applications with fixed phrases or limited vocabularies, such as digits, more explicit pronunciation modeling can be obtained by modeling and comparing acoustic feature sequences over common words or phrases. For text-independent applications, modeling all possible words is not feasible and explicit pronunciation modeling is more difficult. In [1, 2], a system using acoustically adapted automatic speech recognition (ASR) systems for speaker-dependent subword pronunciation modeling was successfully applied to a text-independent task.

In this paper, we present a new technique to explicitly model a speaker's pronunciations by learning a relation between the phonemes the speaker intends to produce and the phones he actually produces. This new approach relies only on the token outputs from a speaker-independent ASR system and open-loop (i.e., no language model) phone recognizers and, thus, is removed from directly using the low-level acoustic signal, which may make it more resistant to signal distortions. When applied to the NIST 2001 Extended Data Task, an equal error rate (EER) of 2.1% was obtained, which further reduces to 0.5% when fused with a GMM acoustic-based system.

2. SYSTEM DESCRIPTION

This section provides an overview of the conditional pronunciation modeling (CPM) system. The aim of this approach is to model speaker-specific pronunciations by learning the relation between what has been said (phonemes) versus how it has been pronounced (phones). For example, a person with a Southern U.S. accent may intend to say 'you' but actually pronounce it as 'yew'. To learn this relation, we rely on the output from an ASR system with its lexical constraints and an unconstrained phone recognizer.

Since we do not know a speaker's intended utterance, we use an ASR system to provide the sequence of intended phonemes. After decoding a spoken word, the possible phonemes for the word from the lexicon (a word can have multiple lexical pronunciations) are force-aligned with the audio and the one with the highest likelihood match score is selected as the phonemic transcription for the word. This provides a time-aligned phoneme stream for all words. For this work, the phoneme sequences from the SRI Prosody Database were used [5].

The phones actually pronounced are taken from an open-loop (i.e., null-grammar language model) phone recognizer. We used phone sequences from five languages: English (EG), German, (GE), Spanish (SP), Japanese (JA), and Mandarin (MA). The phone recognizer is from the Lincoln PPRLM LID system and uses gender-dependent phone models. This output, too, provides a time-aligned phone stream.

For a given utterance, the phoneme and phone sequences are then time aligned at the frame level. An example of these streams and their alignment for the utterance fragment "to you" is shown in the following table:

FRAME	ASR	EG	GE	SP	JA	MA	WORD
24964	t	n	n	n	sh	N	TO
24965	t	s	h	s	sh	N	
24966	t	s	h	s	sh	N	
24967	t	s	h	s	sh	S	
24968	t	s	h	s	sh	S	
24969	t	s	h	s	sh	S	
24970	t	s	h	s	rx	S	
24971	ax	I	h	s	rx	i:	
24972	ax	I	h	iy	rx	i:	
24973	ax	I	h	iy	y	i:	
24974	y	j	h	iy	y	i:	YOU
24975	y	j	i:	iy	y	i:	
24976	y	j	i:	iy	y	i:	
24977	uw	j	i:	iy	y	i:	
24978	uw	u	i:	iy	y	i:	
24979	uw	u	E	iy	y	i:	
24980	uw	u	E	iy	uw	i:	
24981	uw	u	E	sil	uw	i:	
24982	uw	silx	E	sil	uw	i:	
24983	uw	silx	E	sil	sil	i:	

In this table, FRAME is a frame number (at 100 frames per second), WORD is a recognized word, ASR is the phoneme stream, and EG, GE, SP, JA, MA are the language-dependent phone streams.

For a phone stream, EG for instance, we then estimate the conditional probability of EG given ASR (that is, the *realization* given the *intention*) on a per-frame basis. Subsequently, these conditional probability models are used to form a likelihood ratio detector, as described in the following two subsections.

2.1. Training

Since we are using a likelihood ratio detector, we have two models: the background model and the speaker model. Both consist of conditional probabilities of EG given ASR and are trained in the same way:

$$P(\mathrm{EG}=e|\mathrm{ASR}=a)=\frac{\#((e,a)\text{ appears in the INPUT})}{\#((*,a)\text{ appears in the INPUT})}$$

Where $*$ means any EG phone. INPUT is a stream of (EG, ASR) time-aligned pairs, as explained earlier. When training the speaker model, INPUT is from the speaker's enrollment speech. For the background model training, INPUT is from a large amount of speech coming from a speaker set not containing the target speaker. We further assume that INPUT (both in training and testing) has been filtered by removing frames the ASR marks as silence (*sil* phoneme) and frames containing cross-talk (*silx* phone).

In contrast to other phonetic-based speaker recognition systems [6, 8, 7], this approach relies on both the open-loop phone sequences and the constrained phoneme sequences from an ASR system. The conditioning on the phonemes is important because phones alone are quite ambiguous in the sense that a given phone may be caused by many different phonemes (depending on the context, *speaker*, etc.). Conditioning phones on phonemes aims at avoiding this ambiguity. It provides a distribution that carries information about individual pronunciation of phones, as well as some coarticulation habits.

2.2. Testing

The score of a test utterance is computed as follows.

$$\mathit{score} = \sum_{\substack{(e,a)\text{ from INPUT s.t.}\\ \text{both } P_{\mathrm{SP}}(e|a)\text{ and}\\ P_{\mathrm{BG}}(e|a)\text{ are defined}}} (\log(P_{\mathrm{SP}}(e|a)) - \log(P_{\mathrm{BG}}(e|a)))$$

This means that only those pairs (e,a) that have been seen during the training of both the speaker and the background model are counted. P_{SP} represents the speaker model, while P_{BG} is for the background model. Also, we treat INPUT as an array, so, unlike with sets, if there is a pair (e,a) in the INPUT occurring N times, it will be counted N times in the sum.

A conditional probability model and a detector are built for each phone stream and the stream scores are summed to obtain the final system score.

3. EXPERIMENTS

All experiments were conducted on the Switchboard-I (SWB1) corpus according to the NIST Extended Data paradigm [9]. In the NIST Extended Data paradigm, there are 5 training conditions consisting on 1, 2, 4, 8 and 16 enrollment conversation sides, where a side is nominally 2.5 minutes in duration. Testing is done on a entire conversation side. The evaluation consists of a 6-split jack-knife over the entire Switchboard-I corpus. Two background models were used for testing. One model was trained using data from splits 1-3 and applied when testing on splits 4-6; another was trained using splits 4-6 and used for testing on splits 1-3.

The goal of the these experiments was to test whether and how much the conditioning helps the system performance, as well as whether the ASR plays a crucial role or may be replaced by a less computationally expensive step.

The aim of the first experiment was to examine the effect of phoneme conditioning. We constructed two systems. The first uses the conditional probability models as described above. In the second system, we compute the unconditional probability models of the phone streams. To avoid differences due to speech/silence detection, the second system computed counts only over frames where the ASR phoneme stream indicated speech phonemes. We call this unconditioned approach ASR triggered, since the ASR was used to select frames with speech information.

Using combined ASR-triggered scores from streams GE, SP, JA, and MA produced a 7.0% EER on 8 training conversations, whereas GE, SP, JA, and MA, *given ASR*, achieved a 2.7% EER. The raw (nontriggered) unconditional approach (using all frames) achieved a 13% EER. This means that the performance of a recognizer not using ASR should be somewhere between a 7.0 and 13% EER, depending on the quality of its speech activity detector (here, we assume that ASR is the *perfect* speech activity detector and that nonspeech sounds always make the recognition worse, which is likely).

The goal of the second experiment was to test the effect of replacing the ASR stream with a computationally less expensive phone stream. For this, we used the EG stream. In the ASR-triggered setup (where only non-*sil* frames as marked by ASR are used), combined modeling of GE, SP, JA, and MA, *given EG*, achieved a 4.3% EER with 8 training conversations. The raw setup (all frames counted, except crosstalk) achieved a 7.6% EER. This compares to the 2.7% EER of the original system, indicating the phoneme stream is providing information not found in the phone stream.

This result is twofold. First, it shows that the ASR may be replaced by something computationally less expensive than a full-fledged ASR without a dramatic loss of performance. Second, it helps gain insight into the actual working principle of the CPM approach. Along with the idea of modeling purely the speaker-specific phoneme-to-phone mapping, one also has to consider an alternative and not entirely independent view that the phone sequences on a per-frame basis can be seen as vectors of acoustic features with discretized values (in our case, five-dimensional). Since we model these vectors using probabilistic distributions, a similarity to classic GMM-based frame-by-frame modeling offers itself, acknowledging 1) the differences in the feature and function type (discrete nonparametric vs. continuous parametric modeling) and 2) the ASR information is not utilized in a typical GMM text-independent system. The relatively moderate degradation observed in the second experiment may be a supporting factor for the alternative view of the CPM method.

4. ENHANCING THE BASIC SCHEME

Instead of conditioning on the ASR phoneme stream, it would seem more appropriate to condition on some sub-

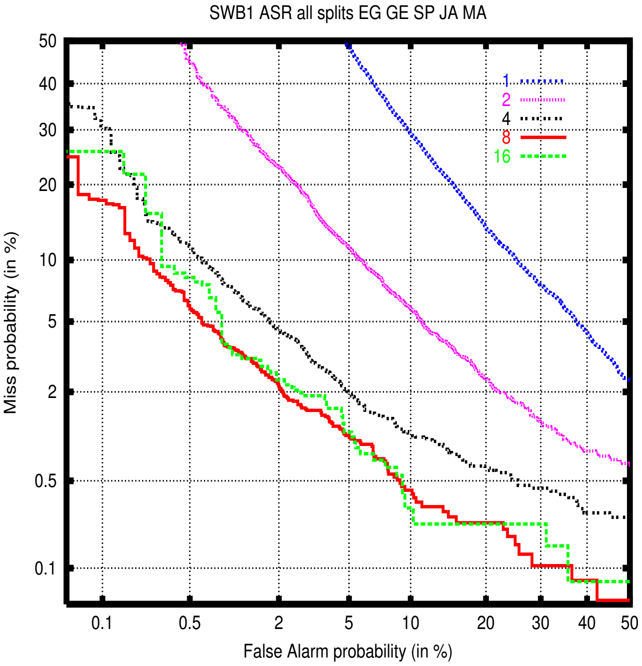

Figure 1. Enhanced system, all streams (2.1% EER for 8-conversation training)

phonemic units; for example, those determined by states of a phoneme Hidden Markov Model. However, we were using a precomputed database with forced alignments not containing such fine segmentation information. Instead, we used an ad hoc solution: assigning states (short/head/body/tail) to the frames of each phoneme according to its length and a frame number counted relative to its beginning. States were assigned in this way: A phoneme shorter than 7 frames is marked as *short*. For phonemes longer than 6 and shorter than 14, the first half of it is marked as *head* and the last half of it is marked as *tail*. For phonemes longer than 13, the 2/7 at the beginning and 2/7 at the end are marked as *head* and *tail*, respectively, leaving the rest as *body*. These marks are then used to define a new phoneme alphabet as a Cartesian product of the original alphabet and the set {*short*, *head*, *body*, *tail*}. The algorithm from section 2. was applied using this enriched alphabet.

On 8-conversation training, this refined modeling gave an improvement to a 2.3% EER (in combined GE, SP, JA, and MA, given the ASR setup, which should be compared with the 2.7% EER from the basic system). When using all five streams of open-loop phones, this approach reached a 2.1% EER. The complete detection error tradeoff (DET) plot of the enhanced-scheme system is shown in Figure 1. Using only the EG stream leads to a 2.8% EER, as shown in Figure 2. Other language-dependent streams did slightly worse than the EG one.

Next, we addressed the question of what result can be obtained using 100% accurate ASR (the ASR in the above experiments had a 30% word error rate). Phoneme alignments obtained from manual word transcripts were available from the SRI database and used in this experiment. All 5 streams achieved a 1.7% EER at 8 conversations, as shown at Figure 3.

System	Alone	Fused with CPM
Acoustics	0.7%	0.5%
Prosodics	6.8%	1.2%
Bintree	3.3%	2.2%
Word n-grams	11%	1.4%

Table 1. Equal-error-rates for fusion of the CPM with other techniques [3]

5. FUSION WITH OTHER SYSTEMS

Fusion experiments were conducted using the enhanced CPM system combined with other systems via a single-layer perceptron, as described in [3]. The CPM system used in the fusion had a 2.3% EER (instead of a 2.1% EER, because it lacked crosstalk (*silx*) filtering). Table 1 summarizes the performance of various fusion configurations (trained on 8 conversations, evaluated on all splits). Performance gains can be seen in all configurations, the strongest relative improvements being achieved by combinations with the prosodics and word n-grams. The latter are also intuitively the least related to the CPM in terms of the information type contained in the features. The least gain is observed with a phonetic modeling technique (Bintree) [8] that exploits statistical dependencies in phone sequences.

6. CONCLUSIONS

The enhanced scheme achieves a relatively high accuracy in comparison with other phonetic methods. It also seems that pronunciation modeling contributes complementary information to both the GMM and other nonphonetic high-level systems. On the other hand, this is at the cost of running the ASR, which makes the method slower than the GMM baseline. But, considering the relatively minor loss in performance when replacing ASR with a phone stream, there may exist other fast predictors to replace the ASR with no performance loss.

Although the results seem to be quite positive, there are a few things that should be mentioned, even though we hope that their effect is negligible. First, the phone recognizers were supplied with (true) external information about the speaker's gender. Second, since the ASR is considered to be a part of our system, the fact that we ran all experiments on Switchboard-I (the same corpus used for training the ASR) means that we mixed training and testing data. Fortunately, the experiment using the EG stream as a predictor suggests that the performance drop should not be too large when the ASR is not trained on test data, even if the ASR worked poorly on data different from Switchboard-I.

7. FUTURE WORK

We observed sensitivity to limited training data. This exhibits itself as poor performance on the 1- and 2-conversation training where there is not enough data to estimate the probabilities properly by raw counting. We plan to use Good-Turing estimates to compute probabilities, which would also make the computation for unseen data less ad hoc than just discarding the data we have not seen during the training.

Second, we would like to investigate the robustness of this method to noise contamination and signal distortions. To enhance it, we want to quantize vectors of phones (coming out of the multiple streams) into a new alphabet on the basis of their noise resistance. If we are really modeling

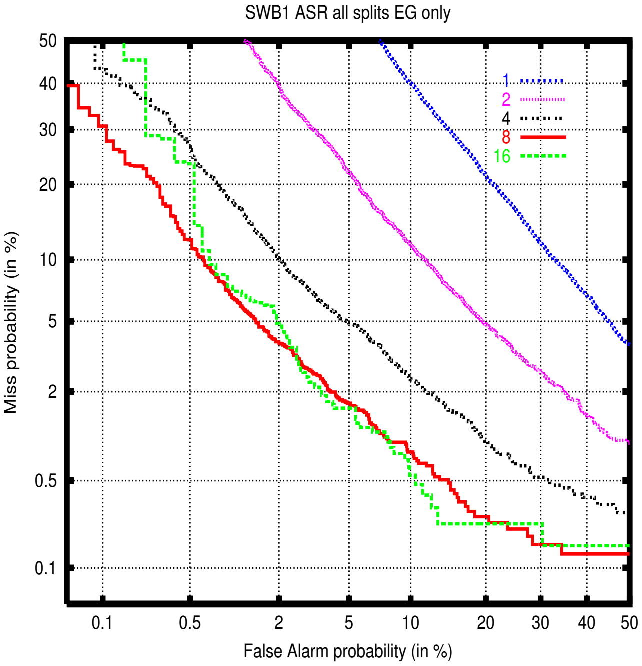

Figure 2. Enhanced system, EG stream only (2.8% EER for 8-conversation training)

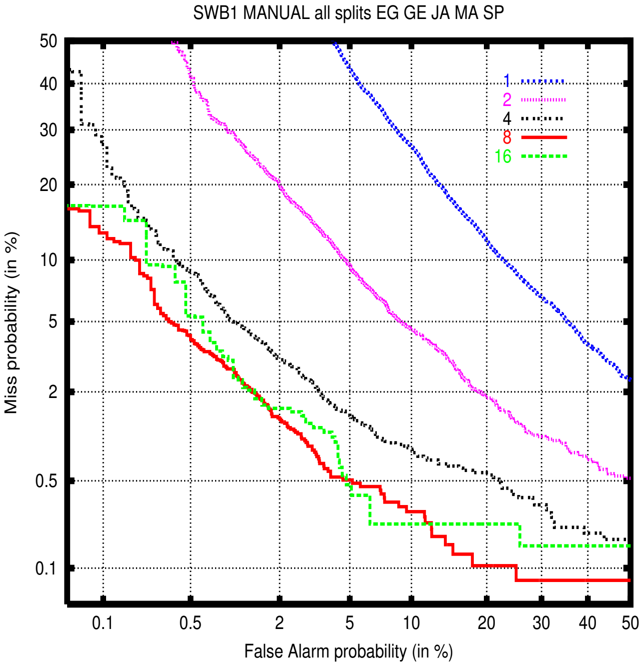

Figure 3. Manual transcript, all streams (1.7% EER for 8-conversation training)

pronunciation, there should be a level of noise or distortion that hides the “color of the voice” recognized by GMM systems, but still leaves pronunciation information in the signal.

Third, since it shows that the ASR may not be necessarily required for the method, we want to investigate the possibility of other predictors; for example, phone recognizers with bigram language models and a suitable speech activity detector.

All of the mentioned modifications should then be validated on new data, particularly on the Switchboard-II telephone speech corpus.

8. ACKNOWLEDGEMENTS

We would like to thank the team of which we were members during the Johns Hopkins University Summer Workshop for their input and encouragement. Also, we would like to thank Professor Frederick Jelinek and the Center for Language and Speech Processing staff for the creative atmosphere they allowed to evolve and for their support.

REFERENCES

[1] M. Newman, L. Gillick, Y. Ito, D. McAllaster, and B. Peskin, “Speaker verification through large vocabulary continuous speech recognition,” In Proc. of ICSLP-96, Philadelphia, PA, 1996, pp. 2419-22.

[2] F. Weber, B. Peskin, M. Newman, A. Corrada-Emmanuel, and L. Gillick, “Speaker recognition on single- and multispeaker data,” In Digital Signal Processing, January/April/July 2000, Vol. 10, No. 1-3.

[3] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, B. Xiang, “The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition,” ICASSP’03.

[4] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, 10(1-3):19–41, January/April/July 2000.

[5] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, ”Prosody-Based Automatic Segmentation of Speech into Sentences and Topics,” Speech Communication, Vol. 32, No. 1-2, pp. 127-154, 2000.

[6] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero, “Gender-Dependent Phonetic Refraction for Speaker Recognition,” In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, May 2002, Vol. 1, p. 149-152.

[7] Q. Jin, J. Navrátil, D. Reynolds, J. Campbell, W. Andrews, and J. Abramson, “Combining cross-stream and time dimensions in phonetic speaker recognition,” ICASSP’03.

[8] J. Navrátil, Q. Jin, W. Andrews, and J. Campbell, “Phonetic speaker recognition using maximum-likelihood binary-decision tree models,” ICASSP’03.

[9] NIST Speaker Recognition 2001 website http://www.nist.gov/speech/tests/spk/2001