

On the Relationship between Phonetic Modeling Precision and Phonetic Speaker Recognition Accuracy

Doroteo Torre Toledano, Carlos Fombella^{}, Joaquin Gonzalez Rodriguez, Luis Hernandez Gomez^{*}*

Área de Tratamiento de Voz y Señal (ATVS)
Escuela Politécnica Superior (EPS)
Universidad Autónoma de Madrid (UAM)
{doroteo.torre,joaquin.gonzalez}@uam.es

^{*}Grupo de Aplicaciones de Procesado de Señal
SSR-ETSIT
Universidad Politécnica de Madrid (UPM)
cfombe@yahoo.com, luis@gaps.ssr.upm.es

Abstract

Speaker recognition techniques have traditionally relied on purely acoustic features and models. During the last few years, however, the field of speaker recognition has started to show interest in the use of higher level features. In particular, phonetic decodings modeled with statistical language models (n-grams) have already shown its effectiveness in several research works. However, the relationship between phonetic modeling precision and the accuracy of phonetic speaker recognition has not yet been sufficiently analyzed. As part of our preparation for the NIST 2005 speaker recognition evaluation, we have performed a number of experiments that show that there is a negligible correlation between phonetic modeling precision and phonetic speaker recognition accuracy. Furthermore, our experimental results show that phonetic speaker recognition results may even be better when using phonetic decodings in languages different from that of the speech.

1. Introduction

Traditional approaches to automatic speaker recognition only consider the acoustic properties of speech, which are usually modeled with statistical models like Gaussian Mixture Models (GMMs). However, recent research has shown that other, higher level, features (e.g. pronunciation idiosyncrasies, linguistic content, prosody, etc.) can also be effectively used in automatic speaker recognition. In particular, numerous experiments have shown that, due to the complementary characteristics of acoustic and higher level features, fusing the information provided by the two kind of features yields further improvements in speaker recognition.

The interest in the use of these higher level features was initiated by the work of Doddington [1]. In his work, Doddington used the lexical content of the speech, modeled using word n-grams, for speaker recognition using the Switchboard-II corpus. This relatively simple technique improved the results obtained by an acoustic-only speaker recognition system.

After the work of Doddington a number of research works have continued exploring the use of higher level features in the field of speaker recognition. One of these works [2,3] made use of similar techniques (n-gram models) applied to the output of open-loop phonetic decoders (i.e. the output of phonetic recognizers without any kind of language model and a grammar consisting of a loop of all the phones). Instead of modeling the lexical content, these techniques aim to model speaker pronunciation idiosyncrasies. The use of this technique gave also very good results, particularly when

several open-loop phonetic decoding for different languages were used and combined. As in the work of Doddington [1], this technique was particularly useful in combination with traditional acoustic-only speaker recognition systems.

After these works, many other researchers have continued analyzing the use of higher level features for speaker recognition [4], and in particular the use of phonetic decodings with n-grams [5,6], but also with other techniques like modeling conditional pronunciations [7] or using binary decision trees [8].

Most of the works using phonetic decodings for speaker recognition have in common that the phonetic decoder is considered as a black box that produces a stream of phones given a spoken input. In this work, however, we are interested in analyzing the relationship between the phonetic modeling precision, the phonetic decoding accuracy and the phonetic speaker recognition accuracy.

The rest of the paper is organized as follows: Section 2 provides a detailed description of the experiments performed. Section 3 presents the phonetic decoding results for the different phonetic decoders tested, and Section 4 presents the speaker recognition results obtained with those phonetic decodings. Finally, Section 5 discusses the results obtained presenting the most important conclusions as well as future research directions.

2. Description of the experiments

The methodology used in this study is based on a very simple idea – instead of using a single phonetic decoder for each different language we used many different phonetic decoders with different complexities (and therefore different phonetic modeling precisions and phonetic decoding accuracies) and compared the speaker recognition accuracies obtained using those phonetic decodings.

The phonetic decoders are based on Hidden Markov Models (HMMs) and implemented using HTK [9]. The phonetic HMMs are three-state left-to-right models with no skips, being the output *pdf* of each state modeled as a weighted mixture of Gaussians. The acoustic processing is based on the Advanced Distributed Speech Recognition Standard Front-End [10], which includes mechanisms for robustness against channel distortion (blind equalization) and additive noise (double Wiener filter).

Phonetic HMMs were trained for Castilian Spanish and American English using the Albayzin [11] and TIMIT [12] corpora. Since these corpora are microphone corpora sampled at 16 kHz, we filtered them to simulate a telephone channel and then downsampled them to 8 kHz before training and testing the models. Only context-independent phonetic

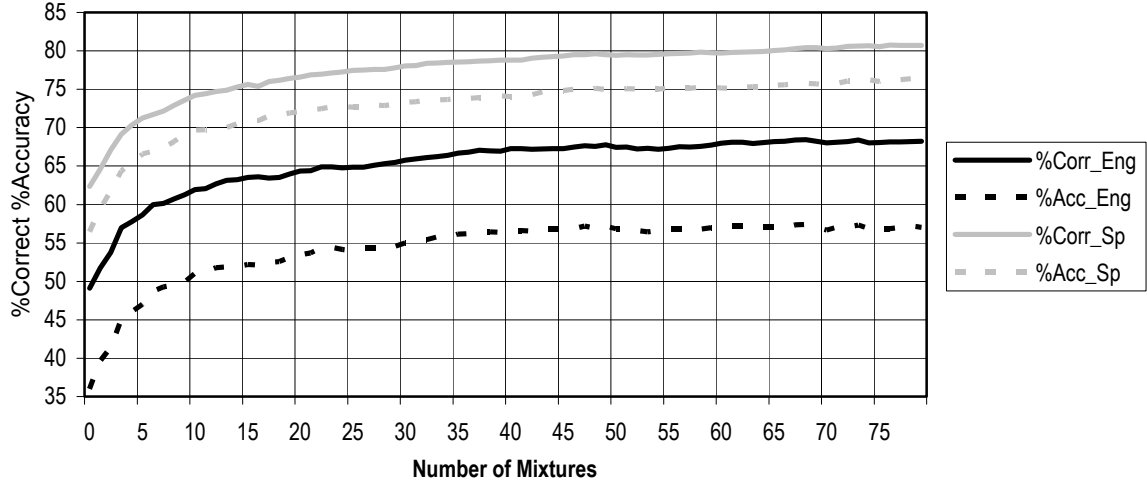


Figure 1: Phonetic decoding results for English (TIMIT) and Spanish (ALBAYZIN) as a function of the number of Gaussians per HMM state. Phonetic decoding accuracy clearly improves with the number of Gaussians, particularly between 1 and 20 Gaussians.

HMMs were used. 23 phones were considered for Castilian Spanish and 39 for American English according to the CMU pronouncing dictionary [13]. To consider different phone model complexities, we trained phonetic HMMs with 1 to 80 Gaussians per state for each of the two languages considered using (part of) the training part of each corpus.

Once the HMMs were trained, we performed phone decodings on a different part of each corpus (not used for training) and evaluated the phonetic decoding accuracy for the different model complexities.

After the phonetic decoding accuracy was evaluated for different phonetic HMM complexities, we performed phone bigram based speaker recognition tests on part of the NIST2002 speaker recognition test corpus and also on the NIST2004 speaker recognition test corpus and analyzed how those results correlate with the phonetic decoding results.

3. Phonetic Decoding Results

Figure 1 represents the evolution of the phonetic decoding results (percentage of phones correctly recognized and phonetic accuracy) for English and Spanish as a function of the number of Gaussians (mixtures) per state in the phonetic HMMs. English acoustic models were trained on the training portion of the TIMIT database. Results for English were obtained on the core test part of TIMIT. For Spanish, models were trained on part of the training portion of the acoustic-phonetic corpus, while testing was performed on a subset containing different speakers.

It can be seen that the acoustic-phonetic accuracy increases as the number of mixtures (and therefore the model complexity) increases. That increase is particularly important between 1 and 20 mixtures, reaching a plateau above 40 mixtures. Results for both languages are almost parallel, the difference being due to the simpler phonemic system of Spanish (we only considered 23 phones for Spanish, while for English we considered 39 phones).

One could expect that the better the quality of the phonetic decoding, the better the quality of the speaker recognition results obtained with those phonetic decodings. Next section will show that this is not the case.

4. Speaker Recognition Results

The phonetic HMMs trained as described in the former section were used to produce phonetic decodings of several standard speaker recognition corpora. These phonetic decodings were used to train phone bigrams for a particular speaker (Speaker Phone Model, SPM_i) and also a global background model (Universal Background Phone Models, $UBPM$). Once the statistical language models were trained, given a test utterance we first produce its phonetic decoding, X , in the same way as the decodings used to train the SPMs and UBPM. Then we use the phonetic decoding of the test sentence, X , and the statistical models (SPM_i , $UBPM$) to compute the likelihoods:

$$\begin{aligned} L_{S_i} &= P(X | SPM_i) \\ L_U &= P(X | UBPM) \end{aligned} \quad (1)$$

Recognition score is the log-likelihood ratio:

$$Score_i = \log \left(\frac{P(X | SPM_i)}{P(X | UBPM)} \right) \quad (2)$$

Using this method, described in [4] as phonetic speaker recognition in the time dimension, we obtained speaker recognition results as a function of the number of Gaussians per phone HMM state for two different corpora.

First, we used a small subset of the NIST 2002 evaluation composed of 15 male and 15 female target speakers (with about 2 minutes of training material for each one), 134 male and 147 female test recordings (with about 30 seconds of speech) for a total of 4215 tests. UBPMs were trained on a different subset of NIST 2002 containing 56 female and 82 male segments ranging from 20 to 120 seconds. With this training and test material we performed speaker recognition tests using 1,5,10,15,...,80 Gaussians per state. Results are shown on Figure 2. These results showed no clear correlation with the phonetic decoding accuracy shown in Figure 1. In

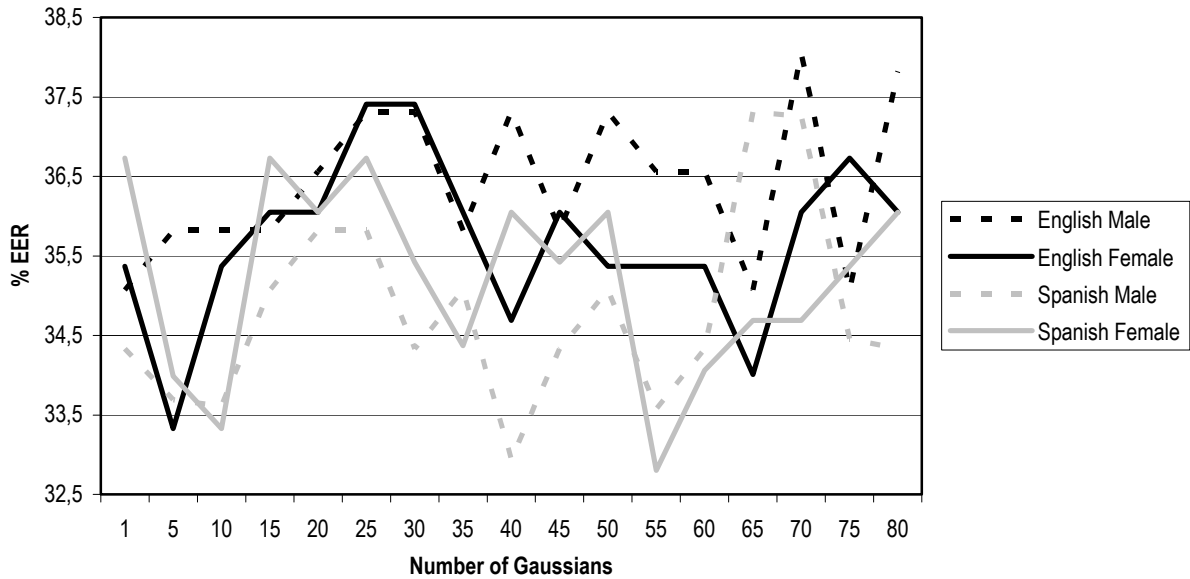


Figure 2: Speaker recognition results expressed as the Equal Error Rate (EER) for a small subset of the NIST 2002 evaluation corpus. Results are shown as a function of the number of Gaussians per phone HMM state.

fact, we computed Pearson's correlation coefficients between phonetic decoding accuracy and Equal Error Rates (EERs) and none of them were statistically significant at the 95% confidence level. We expected the EER to decrease as the precision of the phonetic modeling and therefore the phonetic decoding accuracy increases, but Figure 2 and the results from the correlation analysis do not show any indication of such behavior.

To corroborate this somewhat surprising result, we extended our testing to a more extensive test-bed: the NIST 2004 8sides-1side evaluation, which included 396 target speakers with 8 segments of approximately 5 minutes of speech each, and 16980 different tests using as test speech a segment of approximately 5 minutes of speech. In this case, the UBPM was trained on all the training material available. This can make results a bit optimistic. However, we are interested here in the variation of results with the complexity of the HMMs instead of on the absolute value of the results. Results obtained on this experimental test-bed are shown on Figure 3. In this case, speaker recognition tests were only performed with up to 20 Gaussians/state because the phonetic decoding with more complex models was computationally expensive. In turn we have performed the tests also for 2, 3 and 4 Gaussians per state.

Results obtained are shown in Figure 3 and are only slightly worse than the best phone n-gram based speaker recognition results reported in the NIST 2004 evaluation, around 21% EER fusing the results from several phonetic decodings in different languages. Figure 3 shows that there seems to be a tendency of the EER obtained using English phonetic decodings to increase as phonetic decoding accuracy (and HMM complexity) increases. This tendency is observed using Pearson's correlation coefficient (0.42), but with not enough statistical significance (p -value = 0.47). For Spanish phonetic decodings, however, there is a clear tendency of EER to decrease as phonetic decoding accuracy (and HMM complexity) increases. Pearson's correlation (-0.96) confirms

this statistical significant (p -value = 0.0079) correlation. Best results are obtained using Spanish HMMs with 15 Gaussians per state.

Taking into account that the NIST 2004 evaluation corpus contains around 80% of English test recordings and less than 5% Spanish test recordings, these results might seem surprising. However, this is not the first time that similar results are reported. In [14] a study chose the best combinations of two phonetic decodings (out of 12 possible phonetic decodings in 12 different languages, including English) to perform speaker recognition in English. The best language pairs (Chinese-Korean and Chinese-Spanish) did not include English. Our results also show that best phonetic speaker recognition results are obtained using a phonetic decoding in a language different from that of the test material. Moreover, our results show that while more precise phonetic HMMs in the language of the test material seem to have no effect on (or even worsen) phonetic speaker recognition accuracy, more precise phonetic HMMs in a language different from that of the test material seem to improve speaker recognition accuracy.

5. Conclusions and Future Work

The main conclusion of this research is that there seems to be a negligible, or even negative, correlation between phonetic modeling precision (or phonetic decoding accuracy) and phonetic speaker recognition accuracy attained by modeling the phonetic decodings with n-grams. This conclusion has an important practical consequence: using very simple (even 1-mixture) HMMs for phonetic speaker recognition can be a better (or at least comparable) choice than using very precise and more computational costly HMMs.

This main conclusion is strongly related to the conclusion, also reached in another independent research study [14], that better phonetic speaker recognition accuracy may be attained using phonetic decodings in a language different from that of the test recordings.

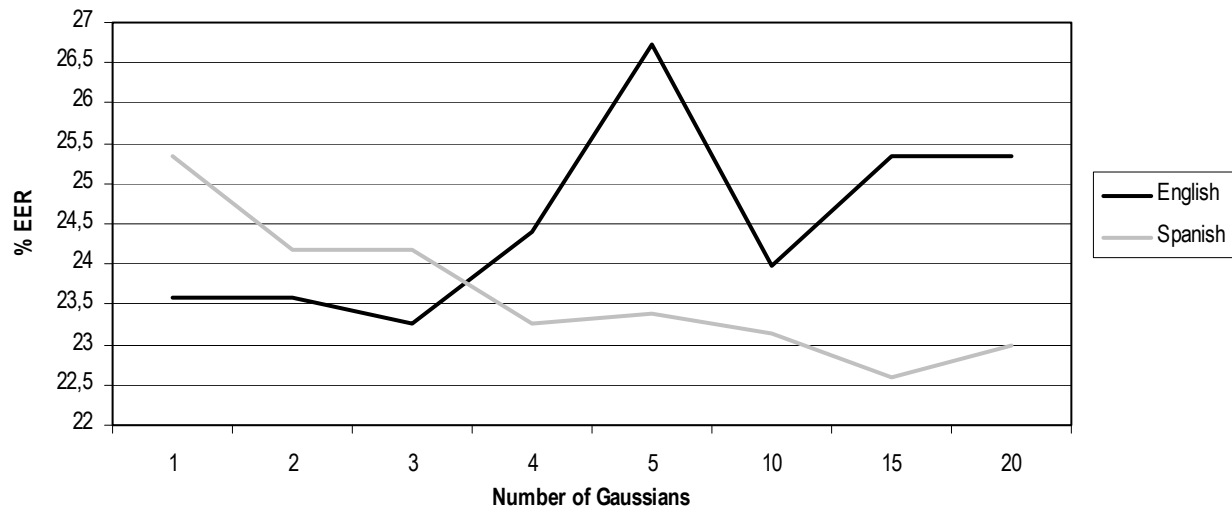


Figure 3: Speaker recognition results expressed as the Equal Error Rate (EER) for the NIST 2004 evaluation corpus (8 sides-1 side condition). Results are shown as a function of the number of Gaussians per phone HMM state.

Our future research will try to confirm these conclusions with further experiments and will also try to answer some very interesting questions raised by these conclusions:

- Are phonetic decoding errors more useful in phonetic speaker recognition than correctly decoded phones?
- How to exploit a more detailed phonetic acoustic modelling for improved phonetic speaker recognition?

6. Acknowledgements

This research was supported by the Spanish Ministry of Science and Technology under project TIC2003-09068-C02-01.

7. References

- [1] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *proceedings of EUROSPEECH*, Vol. 4, pp. 2517-2520, Denmark, 2001.
- [2] W. Andrews, M. Kohler, J. Campbell, and J. Godfrey, "Phonetic, idiolectal, and acoustic speaker recognition," in *Proceedings of ODYSSEY workshop*, 2001.
- [3] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernández-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 1, pp. 149-152.
- [4] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "SuperSID Project Final Report: Exploiting High-Level Information for High-Performance Speaker Recognition", retrieved on march 3, 2005 from http://www.clsp.jhu.edu/ws2002/groups/supersid/SuperSID_Final_Report_CLSP_WS02_2003_10_06.pdf.
- [5] B. Baker, R. Vogt, M. Mason, and S. Sridharan, "Improved Phonetic and lexical Speaker Recognition through MAP Adaptation," in *proceedings of ODYSSEY workshop*, pp. 91-96, Toledo (Spain), 2004.
- [6] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, and J. Abramson, "Combining Cross-Stream and Time Dimensions in Phonetic Speaker Recognition," in *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.
- [7] D. Klusacek, J. Navratil, D. Reynolds, and J. Campbell, "Conditional Pronunciation Modeling in Speaker Detection," in *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.
- [8] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, "Phonetic Speaker Recognition using Maximum-Likelihood Binary-Decision Tree Models," in *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.
- [9] Hidden Markov Model ToolKit (HTK), available on <http://htk.eng.cam.ac.uk/>.
- [10] ETSI ES 202 050 (v1.1.3): "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end features extraction algorithm; Compression algorithms."
- [11] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, C. Nadeu, "ALBAYZÍN Speech Database: Design of the Phonetic Corpus," in *proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH)*. Berlin, Germany, 21-23 September 1993. Vol. 1. pp. 175-178.
- [12] TIMIT "TIMIT Acoustic-Phonetic Continuous Speech Corpus", *National Institute of Standards and Technology Speech Disc 1-1.1*, NTIS Order No. PB91-5050651996, October 1990.
- [13] The CMU Pronouncing Dictionary, available on <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- [14] Q. Jin, T. Schultz and A. Waibel, "Phonetic Speaker Identification," in *proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA, September 2002.