Language dependence in multilingual speaker verification

Neil T. Kleynhans, Etienne Barnard

Human Language Technologies Research Group,
University of Pretoria / Meraka Institute, Pretoria, South Africa
s20109042@tuks.co.za, ebarnard@up.ac.za

Abstract

An investigation into the performance of current speaker verification technology within a multilingual context is presented. Using the Oregon Graduate Institute (OGI) Multi-Language Telephone Speech Corpus (MLTS) database, we found that the performance of text-independent speaker verification depends fairly strongly on the language being spoken, with equal error rates differing by more than a factor of three between the best and worst performing languages. It was also found that training language-specific universal background models, to normalize speakers' scores, gives better results than both language-independent background models and background models derived from relevant language families.

1. Introduction

A speaker verification system needs to determine whether or not a person is indeed who he or she claims to be, based on one or more spoken utterances produced by that individual [1]. A security system based on this ability has great potential in several domains - it is, for example, ideally suited for telecommunications applications since it is non-intrusive, fast, and usable with normal land-line or cellular telephones.

Two forms of speaker verification are typically distinguished, namely text-dependent and text-independent verification [2]. In a text-dependent setup, a predetermined group of words or sentences are used to enroll a set of speakers, and these words or sentences are then used to verify the speakers [1]. In a text-independent system, no constraint is placed on what can be said by the speaker. Text-dependent systems are typically used in combination with pass phrases or personal identification numbers in an explicit verification protocol, whereas text-independent systems generally operate in the background, performing implicit verification while the user is performing other tasks (e.g. talking to an agent or a speech-recognition system). In the current paper, we focus our attention on text-independent verification.

The most popular modelling approach for text-independent systems employs Gaussian mixture models (GMMs) to model the probability densities of acoustic vectors produced by a speaker. This semi-parametric modelling method can represent an arbitrary probability density [3], and can efficiently be calculated and updated as additional data becomes available. In addition, no language-specific information is required for this process; hence, multilingual speaker verification is relatively straightforward compared to the complexities of other multilingual speech-processing systems. An adaptive speaker training method proposed by Reynolds *et al.*, (referred to as a coupled training scheme, since a specific model is adapted from another model) has been shown to outperform a decoupled training method, in which each speaker's model is trained independently [3]. In the coupled training scheme, a speaker's model is obtained by adapting a combination of parameters from a universal background model (UBM). A UBM is a GMM trained with a combination of speakers from either the same database as the test population, or a different database. The adaptation of the UBM parameters is determined by the speaker's data [3] – thus, speaker specific models are created.

In a speaker verification system a threshold (usually a log-likelihood score) is used to either accept or reject a speaker. The value of the threshold can be determined using extra data collected from the speaker during the enrollment phase and can be altered during application of the system, to more closely represent the optimal threshold value of a specific speaker, throughout the use of the system. To improve a verification system's performance, [4] proposed that the log-likelihood score, which results from applying a test utterance to a speaker's model, should be normalized. This is achieved by using a score generated from a subsidiary model, known as the cohort speaker model. A cohort is a selection of speakers whose voice characteristics closely match the target speaker. The cohort model is trained using the selected group's training data. It was found in [4] that as the cohort size increased so did the performance of the verification system.

To date, a large majority of speaker verification systems have been operated in a single-language environment. For use in a highly multilingual context (such as South Africa, India, and much of the developing world), the effect of multiple languages on state-of-the-

art speaker verification systems needs to be investigated. This paper investigates two important issues in multilingual speaker verification, namely

- the dependence of text-independent speaker-verification accuracy on the language spoken, and
- the design of an appropriate UBM for multilingual speaker verification; in particular, whether it is preferable to pool speech from different languages in creating such a model.

The remainder of the paper is organized as follows: section 2 describes the OGI database, section 3 details the verification system, section 4 outlines the experimental setup, data used in the experiments and results obtained. Finally, section 5 discusses the results obtained, and proposes refinements and extensions of this research.

2. MLTS Database

The OGI [5] multi-language telephone speech corpus was used in all our experiments. The data present in the database is telephone quality speech sampled at 8000 Hz. The database consists of speech in eleven languages, namely English, French, Farsi, Hindi, German, Japanese, Korean, Mandarin, Tamil, Spanish and Vietnamese [6]. With the collection of data, each participant was prompted for fixed, region-specific and unconstrained vocabulary speech.

In the fixed vocabulary section, speakers were asked their native language (3s), language spoken most of the time (3s), the days of the week (8s) and numbers zero through ten (10s). Next, in the region-specific section speakers were asked for hometown preferences (10s), hometown climate (10s), occupied room description (12s) and a description of the their most recent meal (10s). Finally, in the unconstrained section each speaker was prompted to talk for one minute on a topic of their choice. This minute of speech was then separated into a fifty and ten second portion. The minute of free speech was not just split into the two portions, as the speakers were warned that the fifty second interval is up and that they must bring their speech to a coherent end [6].

It must be noted that the database is incomplete since many of the individual speakers have missing data recordings.

3. The verification system

The speech signal processing and feature extraction were performed with the HCopy program available as part of the Hidden Markov Model automatic speech recognition toolkit (HTK) [7], and reasonable parameters were chosen based on a combination of experimentation and suggestions in the published literature. A 36 dimensional feature vector was used, made up of 18 mel-frequency cepstral coefficients and their first-order derivatives. The first-order derivatives were approximated over the three previous and successive samples. The coefficients were extracted from a 20-ms frame of speech every 10-ms, with no liftering applied to the resulting coefficients. The filter bank used in deriving the cepstral coefficients consisted of 20 triangular filters and was constrained to a frequency band of 300-3400 Hz. No linear or non-linear channel compensation techniques were applied to the speech signal or resulting coefficients.

The Gaussian mixture model used for both the UBM and speaker model consisted of 512 mixtures and diagonal covariance matrices. The procedure to create the UBM model was to train the speaker models using the speaker's data with the expectation-maximization (EM) algorithm and then to find the average of all these models. The speaker models were created from the UBM model by adapting the mean vectors only and using a relevance factor of 16 (based on results in [3]).

4. Experiments

We first selected a set of languages that were suitable for our cross-language experiments: as mentioned above, the MLTS database is incomplete in that some of the phone records are missing or extremely brief for certain speakers. Therefore, the first task was to find all speakers who had produced a set of preselected recordings. The fifty second free speech recording was chosen for training the speaker models; the remaining ten second free speech segment was used for testing, while the ten second hometown climate recordings were used for determining the speaker's threshold scores.

Additionally, it was observed that the two shorter recordings chosen for our experiments, which are each nominally 10 seconds in duration, were actually much shorter for many speakers; thus, further speakers were removed from the database when a minimum duration limit of five seconds for each recording was not met. Eventually, a total of 370 speakers remained in the experimental database. The different languages and speaker numbers that made up the experiment database are summarized in Table 1.

As can be seen in Table 1 the Farsi, Hindi, Korean and Vietnamese languages were removed from the experiment since an insufficient number of speakers remained in each of these languages.

The training termination criteria for both the EM and adaptive training algorithms were chosen as follows: when the value of the log-likelihood score went below 10% of the previous score's value or if 30 training iterations were reached, training was terminated.

Language	Female	Male
English	30	76
French	9	26
German	23	33
Japanese	16	30
Mandarin	15	20
Spanish	13	38
Tamil	5	36

Table 1: Languages from OGI MLTS that were used for cross-language experiments, and the number of speakers (female and male) per language.

4.1. Experiment 1

For the first experiment, both language-specific UBMs and all-language UBMs were trained and used to normalize individual speakers' scores. The plots in figure 1 show the DET curves [8] obtained for the seven languages in our corpus; the legend indicates the language tested and the language(s) used to train the UBM. The keyword "entire" means that all the language data was used to train the UBM, and the Spanish results are repeated in all three graphs in order to provide a basis for comparison.

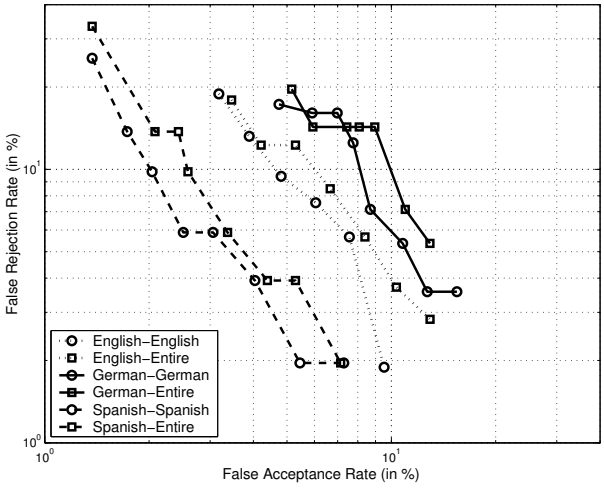
As can be seen in figure 1, the plots are somewhat irregular, as is to be expected from the limited number of test speakers per language. Spanish and Tamil were the best performing languages, with equal error rates around 3%. For French, in contrast, the measured equal error rate was almost 10%, and for German around 8%.

The other salient fact in these figures is that all languages, with the exception of Tamil, perform better when a language-specific UBM is employed. This phenomenon was explored in further detail in Experiment 2.

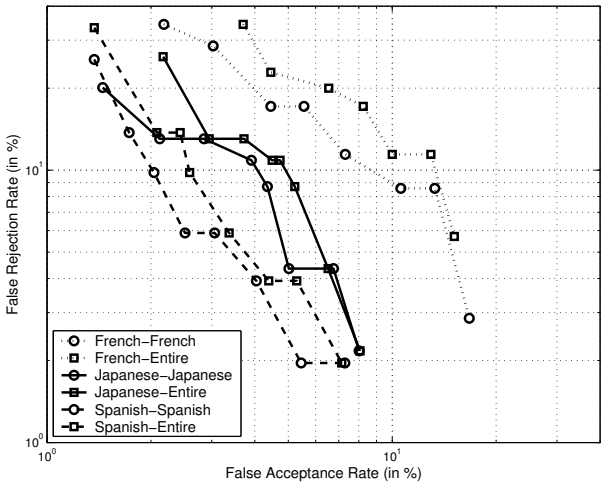
4.2. Experiment 2

In this experiment, we wanted to see whether UBMs intermediate to both the language-specific and language-independent models could be found with better performance than either. Thus, two "language-group" UBMs were trained, namely a Germanic UBM (using English and German data), and a Romance UBM (using French and Spanish data). The two UBMs that resulted were then used to normalize the languages involved in their creation.

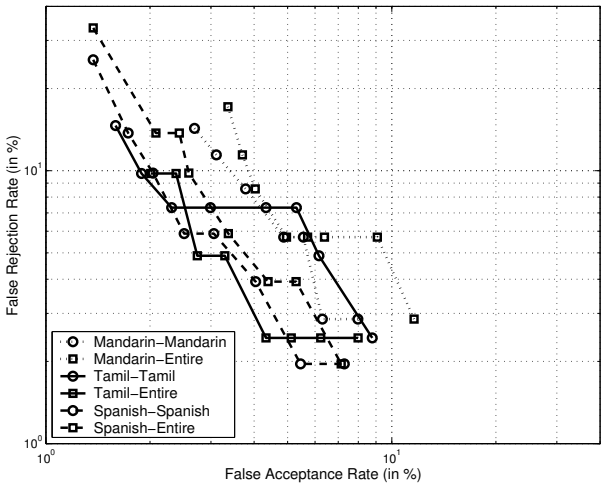
In both figures 2 and 3, the DET curves indicate that the Germanic and Romance UBMs were a better choice over the UBM trained with all the data. The Germanic and Romance UBMs show comparable results in comparison with the same-language UBM, but are consistently somewhat inferior.



(a)



(b)



(c)

Figure 1: The DET curves for different languages using various UBMs. The UBMs were either trained with data from the target language, or with the entire database (ENTIRE).

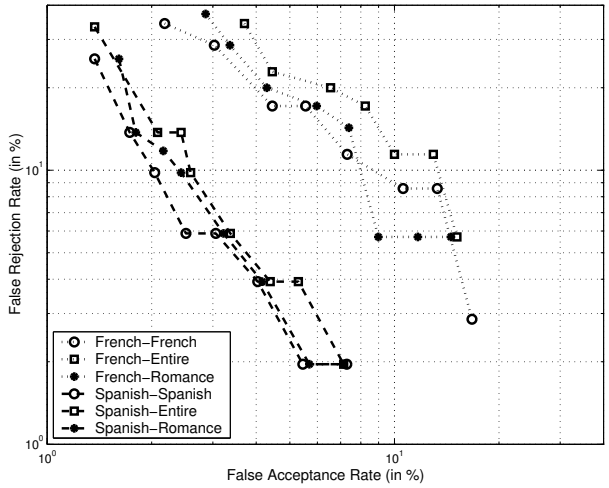


Figure 2: DET curves for French and Spanish languages using different UBMs. The UBMs were trained with a specific language's data, with the entire database (ENTIRE) or the Romance language-group data.

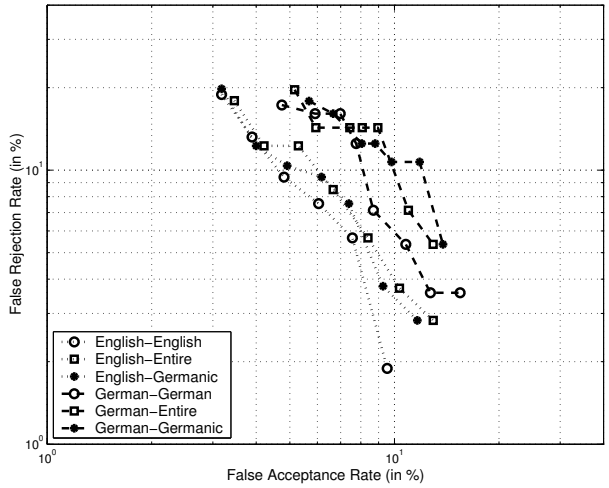


Figure 3: DET curves for English and German languages using different UBMs.The UBMs were trained with a specific language's data, with the entire database (ENTIRE) or the Germanic language-group data.

5. Conclusion

We find substantial differences in the speaker-verification accuracies obtained with different languages. These differences do not correlate with the number of training speakers or the total duration of available speech, which suggests that these are real inter-language differences. The existence of such differences is not unexpected in light of the differences in phonetic content, phonotactic constraints, speaking rhythms, etc. that exist amongst the different languages. However, the magnitude of the observed language differences is quite surprising.

It is also consistently seen that language-specific UBMs lead to improved verification over more general background models (i.e. those trained with data within a language family, or across all data). This is interesting in light of the fact that fairly limited training data was available, so that language-independent UBMs may have been expected to benefit from the additional training data. We therefore expect that this benefit of language-specific UBMs will be even more pronounced in the presence of larger training corpora.

All of the experiments in this paper were carried out with the OGI MLTS corpus, which was originally not designed for the purpose of speaker verification. It would be interesting to repeat these measurements on other corpora with larger numbers of speakers per language, in order to assess how robust these results are. It would also be interesting to systematically investigate other language families, and to understand the factors that determine speaker-verification accuracy in a given language.

6. References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1–26, September 1997.

[2] R. Auckenthaler, M. J. Carey, and J. S. D. Mason, "Language dependency in text-independent speaker verification," in *ICASSP 2001*, May 2001, pp. 441–444.

[3] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[4] C-S Lui, H-C Wang, and C-H Lee, "Speaker verification using normalized log-likelihood score," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 56–60, January 1996.

[5] "Oregon Graduate Institute," 10 September 2005, `http://cslu.cse.ogi.edu/`.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of the International Conference on Spoken Language Processing*, October 1992, pp. 1895–1898.

[7] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, "The HTK book. revised for HTK version 3.3," September 2005, `http://htk.eng.cam.ac.uk/`.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997, pp. 1895–1898.