LANGUAGE DEPENDENCY IN TEXT-INDEPENDENT SPEAKER VERIFICATION

R. Auckenthaler[1,2], M. J. Carey[1], J. S. D. Mason[2]

[1]Ensigma Technologies. Turing House, Station Road, Chepstow, NP16 5PB, UK.
[2]Department of Electrical & Electronic Engineering, University of Wales, Swansea, SA2 8PP, UK

{Roland.Auckenthaler, Michael.Carey}@ensigma.com, J.S.D.Mason@swansea.ac.uk

ABSTRACT

Applying speech technology in appliances available around the world cannot restrict the functionality to a certain language. However, most of today's text-independent verification systems based on Gaussian mixture models, GMMs, use an adaptive approach for training the speaker model. This assumes that the world model incorporates the same language as that of the target speaker.

In this paper we investigate language mismatches between the target speaker and the world model in a GMM speaker verification system. Experiments performed with different world model languages showed major degradations, in particular for Mandarin and Vietnamese when the target speakers spoke American English. Experiments with world models trained on data pooled from different languages revealed only minor performance degradations.

1. INTRODUCTION

With today's technologies in the wireless communication sector the use of mobile devices is not restricted to one country. The same mobile phone can be used all over Europe and in future around the world. This global market requires that speech technology in such devices is not restricted to a particular language. In the specific task of speaker verification multi-lingual use can cause difficulties and may reduce the verification performance.

The language problem may not be important in text-dependent speaker verification applications [1]. A password phrase can be chosen by the user in any language. Text-dependent speaker verification systems create templates for storing the phrase and therefore are able to process phrases from different languages.

In text-independent speaker verification systems the task is slightly different. Gaussian mixture models, GMMs [2], are the preferred was of modelling the speech. The mixture components do not represent a specific password phrase but the voice characteristics of a specific person. The speech used to train these models is limited to a small amount. A world model [3] is used in part to combat the lack of training data. This world model describes the global speech space and is trained using a large amount of data. The speaker model is then adapted from the world model. In work published to date the language of the world and speaker models have been the same. Therefore such a speaker verification system may have potential short falls when a speaker uses a language not used in the creation of the world model.

A major problem of carrying out investigations into language mismatches is that no speaker recognition databases exist including target speakers from different languages or speakers from a multi-lingual background. The lack of databases can be overcome when the principal problem is reversed. An experimental setting is proposed using speakers from one specific language. The speaker verification system incorporates world models from different languages. The world models can be trained for each individual language using databases for language identification tasks [4]. These databases have the advantage that the recordings were performed using the same environment and therefore the same mismatch between each individual language and the speakers exist. The use of several languages also allows training one world model with data pooled from all languages.

In this paper we concentrate on world model training with different languages to perform speaker verification experiments. In the next section the databases are explained in some details. Section 3 describes the speaker verification system. Experimental results are given in Section 4 and Section 5 concludes the experiments and gives suggestions for further work.

2. DATABASES

Two different databases were used for the experiments. The NIST 1998 speaker recognition evaluation database was used for training the speaker models and the NIST

1996 language identification development database was used for world model training.

The NIST 1998 database is part of the switchboard corpus [5] and consists of 200 speakers of each gender speaking American English. Each speaker recorded two sessions for training purposes. Several sessions were recorded for testing. Utterances were taken from each test session with three, ten and thirty seconds in length.

The NIST 1996 language identification development database is part of the Callfriend database [6] and contains speech material from twelve different languages. These are

1. Arabic
2. American English
3. Farsi
4. French
5. German
6. Hindi
7. Japanese
8. Korean
9. Mandarin
10. Spanish
11. Tamil
12. Vietnamese

About 35 minutes of speech were available for training each language. Recordings were performed by an unknown number of speakers in their native language. Both databases are telephone quality sampled at 8kHz.

2. VERIFICATION SYSTEM

Feature vectors were extracted using an LPC front-end with a frame rate of 20ms and an adaptive energy detector. This front-end is part of the enhanced full rate speech codec used in the GSM mobile telephony standard [7]. The LPC features were transformed to cepstral coefficients. Channel normalisation was performed in the cepstral domain by RASTA filtering [8]. First order derivatives were calculated from the channel normalised features. The resulting feature vectors of order 21 consisted of ten static and ten dynamic and the first order derivative of the RASTA filtered frame energy.

The statistical modelling used adaptive GMMs [2]. The world model was trained on the Callfriend corpus whereas the speaker models were adapted from the world model using the switchboard database. Only the mean parameters of the mixture components were used for the speaker model [9]. Speaker models were trained using two minutes of speech from two recording sessions, one minute from each session. Tests were performed using utterances of ten seconds from the same corpus. Score normalisation was performed by subtracting the world model score from the speaker model score. More sophisticated normalisation techniques [10] were not used to avoid effects from cohort speakers, which introduce the possibility for another language mismatch between target and cohort speakers.

2.1 World Model Training

The major issue discussed in this paper is the world model training. Two approaches were used. The first approach trained models for each individual language separately. The twelve world models were then used in turn for the speaker adaptation and the verification experiments.

The second approach for world model training was to pool all the training data from the different languages together. A model was trained with all data, another model was trained with a twelfth of the data from each language. The latter model is a straight comparison to the individual language world models. Pooled models were trained for 64 and 256 mixture components due to the larger amount of training data when pooling all languages together.

3. EXPERIMENTS

The experiments were based on test utterances where the target speaker used the same phone number for training and testing; the same number condition.

The first experiments were conducted using the individual language world models with 64 mixture components. The DET plots [11] in Figure 1 show the individual languages in comparison to American English. This language is the baseline due to no language mismatch between world and speaker models. Large performance degradations were obtained for a Vietnamese and Mandarin world model followed by Arabic, French, Tamil and Spanish. World models trained on German, Hindi, Japanese and Korean revealed similar performances compared to American English. Surprisingly, the Farsi world model showed a slight improvement over the baseline. The reason for this effect is unknown.

The second set of experiments was performed using the multi-lingual world models for model sizes of 64 and 256 components. Figure 2 shows the performance characteristics for both model sizes. A degradation in performance is seen when the world model was trained with all languages and the same amount of training data as the individual American English model. The performance of the multi-lingual world model is similar to the American English when using all available data for training a multi-lingual model. A slight degradation was obtained for the

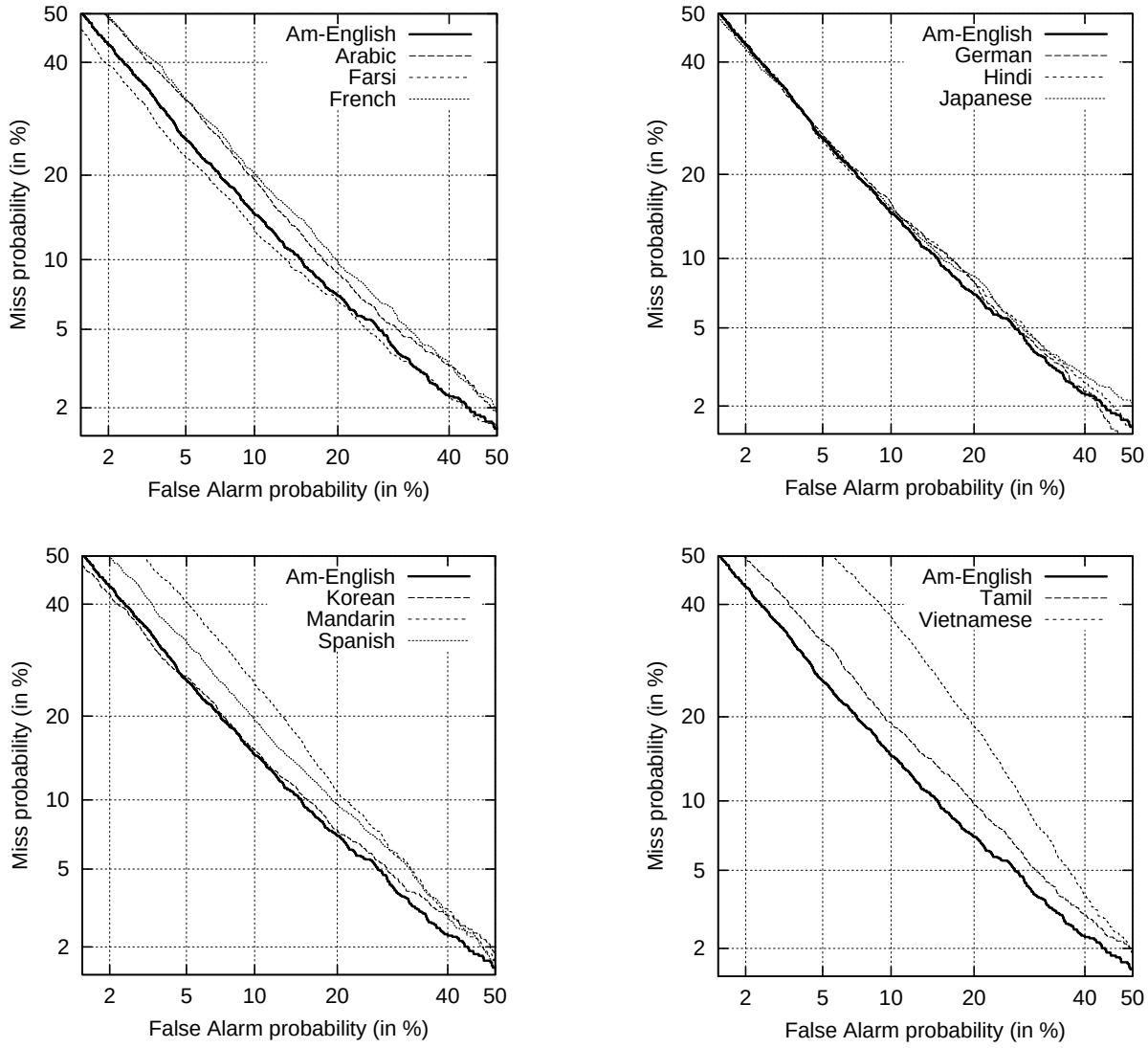


Figure 1: Verification Performances for Individual Language World Models

multi-lingual world model trained with all data and a size of 256 components. This may be due to insufficient training data.

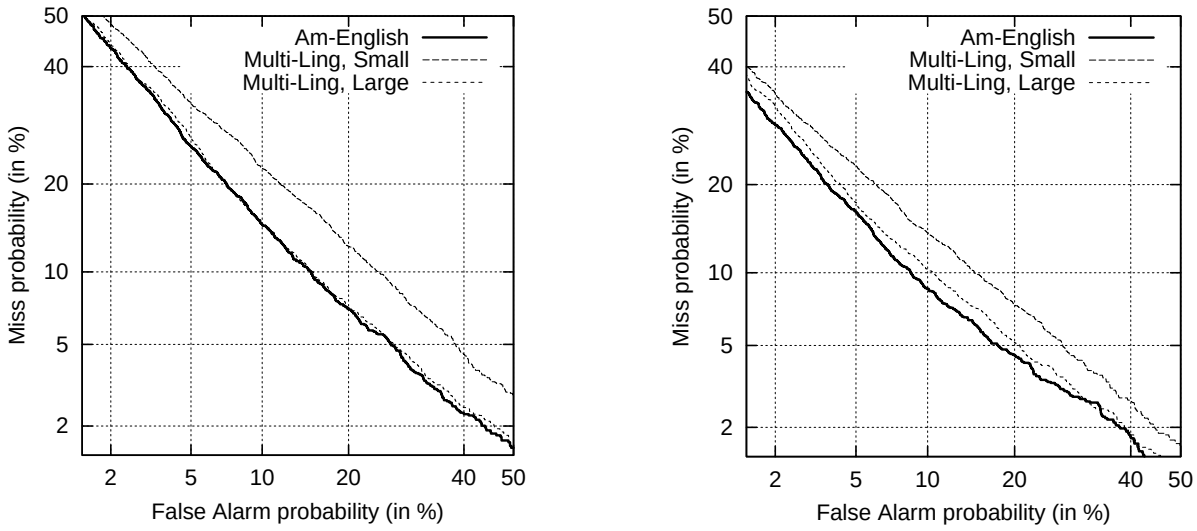
4. CONCLUSIONS

In this paper the impact of language mismatches is discussed for text-independent speaker verification systems. Language is an issue when using an adaptive verification system incorporating a world model for bootstrapping the adaptation process. The world model and the target speaker may not necessarily be from the same language.

Experiments were carried out using different languages to train the world model whereas the target speakers were taken from the switchboard corpus, a database in American English. World models were also trained with a pool of training data covering several languages.

Results revealed large degradations for some languages, in particular for Mandarin and Vietnamese. Some languages performed similarly to American English. Test with multi-lingual world models obtained no performance degradation when enough data were available for world model training.

So far the work carried out for this paper was focused on one target speaker language only due to the lack of



a) 64 Mixture Model

b) 256 Mixture Model

Figure 2: Verification Performance of Multi-Lingual World Models

databases. More work must be carried out to obtain verification results of systems designed for several languages. For this purpose new databases must be collected encompassing speakers from different language backgrounds. This will lead to a better understanding of the impact of language on speaker verification performance.

5. REFERENCES

[1] O. Siohan, C. Lee, A. Surendran and Q. Li, *Background Model Design for Flexible and Portable Speaker Verification Systems*, in Proc. ICASSP, vol. 2,pp. 825-828, Arizona, 1999

[2] D. Reynolds, *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*, in Speech Communication 17:91-108, August 1995

[3] M. Carey, E. Parris and J. Bridle, *A Speaker Verification System using Alpha-Nets*, in Proc. ICASSP, pp. 397-400, Toronto 1991

[4] NIST, *The NIST 1996 Language Recognition Evaluation Evaluation*, http://www.nist.gov/speech

[5] J. Godfrey, E. Holliman and J. McDaniel, *SWITCHBOARD: Telephone Speech Corpus for Research and Development*, in Proc. ICASSP, pp. 517-520, San Fransisco, 1992

[6] The Linguistic Data Consortium, LDC, http://www.ldc.upenn.edu

[7] TIA/EIA Interim Standard, *TDMA Cellular/PCS - Radio Interface - Enhanced Full-Rate Speech Codec, TIA/EIA/IS-641*, May 1996

[8] H. Hermansky and N. Morgan, *RASTA Processing of Speech*, in IEEE Transaction on Speech and Audio Processing, no. 4, vol. 2, pp.578-589, IEEE, October 1994

[9] D. Reynolds, T. Quatieri and R. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models*, in Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, Academic Press, 2000

[10] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, *Score Normalisation in a Text-independent Speaker Verification System*, in Digital Signal Processing, vol. 10, no. 1, Academic Press, January 2000

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, *The DET Curve in Assessment of Detection Task Performance*, in Proc. Eurospeech, pp. 1895-1898, Rhodes 1997