CROSS-SESSION VARIABILITY COMPENSATION IN LANGUAGE DETECTION TASK

Valiantsina Hubeika
Master Degree Programme (2), FIT BUT
E-mail: xhubei00@stud.fit.vutbr.cz

Supervised by: Pavel Matejka
E-mail: matejka@fit.vutbr.cz

ABSTRACT

Presently, demand on automatized language recognition systems is growing rapidly. Variety of recognition systems have been developed by a number of research laboratories using techniques based on different approaches. However these methods vary essentially in their principles, they share same intends. First, it is necessary to capture intra-language variability in order to perform better separation between languages. Secondly, the system should be capable of compensation on so-called cross-session variability. The paper deals with cross-session variability compensation in the acoustic language identification system. Two approached are presented and their performance is analyzed. The presented system was built as a contribution into the submission of Brno University Technology to the Language Recognition Evaluation (LRE) 2007 organized by National Institute of Standards and Technology (NIST).

1 INTRODUCTION

The goal of language recognition is to recognize a language from speech (should not be confused with speech recognition). The main reason of language recognition systems development is its application in international call-centers, where the language has to be recognized from a short segment of speech in order to be switched the call to an operator with appropriate knowledge of the language.

This work concentrates on building-up an acoustic recognition system robust against channel distortion. An acoustic system models the distribution of features reflecting the 'sound' of the speech. To compensate on channel variability so-called eigenchannel adaptation in model domain [1] is primarily used. The approach although has a disadvantage: it is applicable only on a specific already built model, Gaussian Mixture Models (GMM). To transcend this limitation, an approximation of the method was developed, so-called eigenchannel adaptation in feature domain. Both methods are involved in this work, brief theoretical background is given, several experiments consequently presented and the results are compared with other systems.

The organization of the paper is as follows: section 2 introduces main structure of the acoustic system; section 3 deals with eigenchannel adaptation in model and feature domain; experiments can be found in section 3; conclusions and future work is in section 4.

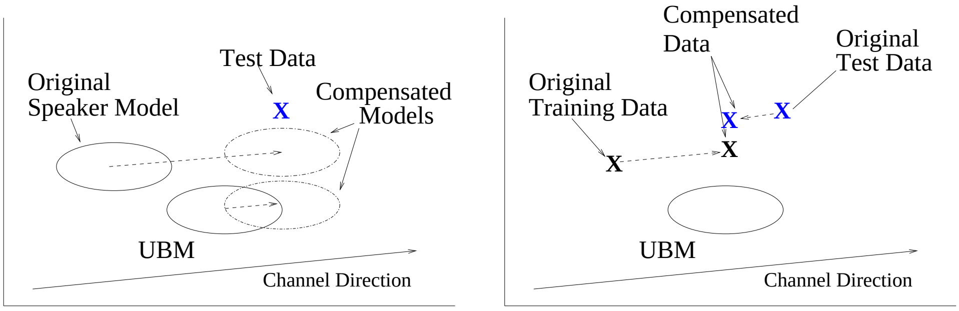

Figure 1: Eigenchannel adaptation in both model (left) and feature (right) domain.

2 ACOUSTIC LANGUAGE RECOGNITION

The recognition system is composed of 2 main parts: front-end and classification. Front-end includes feature extraction, Shifted Delta Cepstra (SDC) [2], and consequent feature transformation, Vocal Tract Length Normalization (VTLN) [2]. This results in a sequence of 56 dimensional feature vectors with shift of 10 ms. The distribution of the features is then modeled using Gaussian Mixture Models (GMM), where each Gaussian is represented by mean vector and diagonal covariance matrix. The approach, called UBM-GMM (UBM stands for Universal Backgroun Model) is used. The UBM is a language independent model with 2048 Gaussians. Language dependent models are MAP adapted from UBM (adapting mean vectors) using language dependent (enrollment) data.

3 CROSS-SESSION VARIABILITY

Often enrollment and test data are recorded over different channels or under different conditions which lowers the accuracy of correct recognition significantly. Under such conditions, it can easily happen that the language will be recognized according to the session conditions (cell-phone vs. land-line) in the recording and not by the language-related information. The main idea of eigenchannel adaptation technique is to find directions in the space of model parameters that represent changes in channel and remove it.

3.1 EIGENCHANNEL SUBSPACE ESTIMATION

Let supervector be a vector of concatenated GMM mean values divided by corresponding standard deviation. Before eigenchannel adaptation can be applied, the directions in which the supervector is mostly affected by a changing channel must be identified. These directions, referred to as eigenchannels, are eigenvectors of average within class covariance matrix, where each class is represented by supervectors estimated on different segments of the same language.

3.2 EIGENCHANNEL ADAPTATION IN MODEL AND FEATURE DOMAIN

The basic principle of eigenchannel adaptation in both model and feature domain is depicted in figure 1. In model domain eigenchannel adaptation is applied on a language supervector during testing. The supervector is shifted in the channel directions to better fit the test data (fig. 1, left).During eigenchannel adaptation in feature domain, feature vectors are shifted to better fit

Table 1: *Results for the presented system and two counterpart systems (right).*

	Baseline	CC-MD	CC-FD	CC-FD, MMI	GMM-SVM-512	11 HU Tree
30 sec	8.03	2.76	2.77	2.41	3.69	3.86
10 sec	12.89	8.64	7.21	6.42	8.77	9.45
3 sec	21.77	19.04	18.08	16.58	20.90	20.37

the UBM supervector (fig. 1, right). Here, both training and test data are transformed. Detailed description can be found in [1].

4 EXPERIMENTS

The language recognition system comprised 14 languages. The amount of the training data for each language varied (from 1.4 to 264 hours). Test data with nominal length of 30, 10 and 3 seconds were used. The results are presented in terms of a standard metric in detection task, Equal Error Rate (EER). Experiments with no channel compensation (Baseline), with channel compensation in model (CC-MD) and feature (CC-FD) domain are shown in the table 1. Eigenchannel adaptation in feature domain slightly outperforms the adaptation in model domain, mainly for the short-duration segments. Additionally, compensation on features brings further advantages as different types of models can be trained (here, Maximum Mutual Information training technique [2]). Nevertheless, both methods bring a dramatic decrease of the error against baseline.

Table 1 presents the accuracy of the best single language recognition subsystems contributed to the final BUT system submitted for LRE 2007. GMM-SVM-512 subsystem [2] is a SVM classifier where the features are GMM supervectors (512 Gaussians). Each language class is defined by a set of supervectors, each trained on one speech segment. 11 HU Tree is a decision tree based language model [3].

5 CONCLUSION

The system with eigenchannel adaptation in feature domain was the best performing subsystem from the BUT subsystems developed for LRE 2007. The result clearly show that the method is efficient and further investigation may bring additional improvement.

REFERENCES

[1] Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., Laface, P.: Channel Factors Compensation in Model and Feature Domain for Speaker Recognition, Speaker and Language Recognition Workshop, IEEE Odyssey, San Juan, PR, 2006,

[2] Černocký, J., Matějka, P., Burget, L., Schwartz, P.: Brno University of Technology System for NIST 2005 Language Recognition Evaluation, Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop, San Juan, PR, 2006

[3] Navratil, J., Recent advances in phonotactic language recognition using binary-decision trees, Interspeech, 2006.