UEF-NTNU System Description for Albayzin 2010 Language Recognition Evaluation

R. Saeidi[1], *M. Soufifar*[2], *T. Kinnunen*[1], *T. Svendsen*[2], *and P. Fränti*[1]

[1] School of Computing, University of Eastern Finland, Joensuu, Finland
[2] Department of Electronics and Telecommunications, NTNU, Trondheim, Norway
{rahim,tkinnu,franti}@cs.joensuu.fi, {soufifar,torbjorn}@iet.ntnu.no

Abstract

We are describing University of Eastern Finland and Norwegian University of Science and Technology joint submission for Albayzin 2010 language recognition evaluation. We are employed several several approaches including acoustic and phonotactic based algorithms in our final submission. A short description of the systems are given.

Index Terms: Language Recognition, GLDS, GMM, MMI, VSM.

1. Submission Overview

Our submission for Albayzin 2010 is a score-level fusion of 3 sub-systems as follows:

- MMI-GMM
- GLDS-SVM-NN
- HMM-VSM-GMM
- PR-VSM-GMM

2. MMI-GMM

This system is build based on [1] and specifications are:

- SDC features of 49 dimension extracted.
- Language dependent 256 Gaussian GMMs are used for modeling.

3. GLDS

This system is build based on [2] and specifications are:

- SDC features of 49 dimension extracted.
- Up to 3rd order polynomial expansion used.
- Neural network with one hidden layer applied for language classifier. Other specifications are: Activation function tansig for hidden nodes and purelin for output layer, MSE measure and trained with trainlm (I used MATLAB terminology here).

4. PR-VSM-GMM

This system is build based on [3] and specifications are:

- Brno university phone recognizers [4] used here. There are 4 phone recognizers in their website. English phone recognizer trained on 16kHz TIMIT data which we used it for Albayzin evaluation.
- Up to 2-gram counts used and 300 dimensions retained after SVD.
- Two GMMs trained for each language in language classifier stage; one GMM for target scores (with 2 Gaussians) and another for non-target scores (with 20 Gaussians).

5. HMM-VSM-GMM

This system is build based on [3] and specifications are:

- Train a UBM on all languages data with 128 Gaussian.
- Tokenize the same data with UBM.
- Using the labeled data train a HMM.
- Treat the HMM as an *event recognizer* and proceed with VSM back-end.

6. Tasks

The task is defined to be language detection for 30s-30s train-test in closed-set detection. Performance measure is Equal Error Rate (EER) and cost function C_{avg} defined by NIST [4].

6.1. Albayzin 2010

- Six languages: Spanish, Catalan, Basque, Galician, Portuguese and English.
- Speech data are extracted from multi-speaker TV broadcast recordings.
- Almost 10 hours of data per language is available for training. There are also some extra noisy data for training systems to deal with noisy situation.
- 836 test samples of 30s length for development set.
- 4992 test samples of different lengths (3, 10 and 30s) for evaluation. We should report our results without considering the length (or the state of being clean or noisy) of the utterance.

7. References

[1] Matejka Pavel, Burget Lukas, Schwarz Petr, Cernocky Jan: Brno University of Technology System for NIST 2005 Language Recognition Evaluation, In: Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop, San Juan, PR, 2006, p. 57-64.

[2] Campbell, W., Campbell, J., Reynolds, D. A., Singer, E., Torres-Carrasquillo, P., Support Vector Machines for Speaker and Language Recognition, Computer Speech and Language, Vol. 20, No. 23, pp. 210229, April 2006.

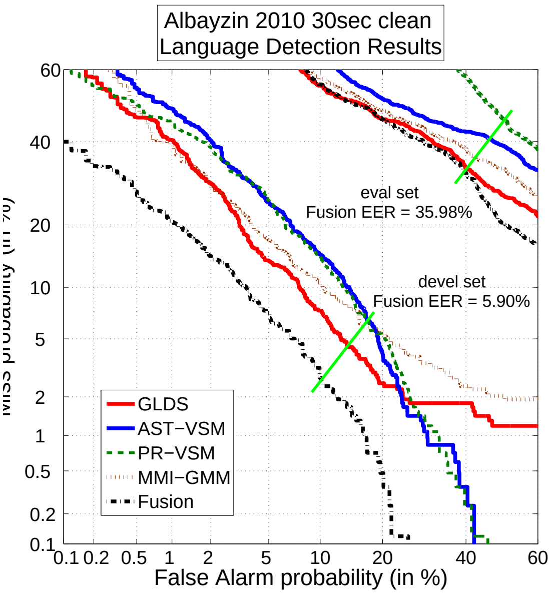

Figure 1: Submitted system result on 30sec clean data.

[3] Haizhou Li; Bin Ma; Chin-Hui Lee; , ”A Vector Space Modeling Approach to Spoken Language Identification,” Audio, Speech, and Language Processing, IEEE Transactions on , vol.15, no.1, pp.271-284, Jan. 2007.

[4] P. Schwarz, ”Phoneme Recognition based on Long Temporal Context, PhD Thesis”, Brno University of Technology, 2009.

[4] www.itl.nist.gov/iad/mig//tests/lre