University of the Basque Country + Ikerlan System for NIST 2007 Language Recognition Evaluation

Mikel Penagarikano[1], German Bordel[1], Luis Javier Rodriguez[1], Juan Pedro Uribe[2]

(1) Department of Electricity and Electronics, University of the Basque Country, Spain
(2) Ikerlan - Technological Research Center, Spain
E-mail: mikel.penagarikano@ehu.es

1. Introduction

This paper briefly describes the language recognition system developed by the GTTS group at University of the Basque Country in collaboration with IKERLAN Technological Research Center, and submitted to the NIST 2007 Language Recognition Evaluation. The system does not use any phonetic, phonological nor morphosyntactical knowledge about the target languages, and is based on simple GMM tokenizers and token n-grams estimated through Maximum Mutual Information (MMI). A couple of issues related to using the MMI criterion to estimate model parameters are addressed, which may give a new insight into this task.[1]

2. System EHUIKER

2.1. System description

The system first performs GMM tokenization; then the resulting sequence of tokens is scored by language-specific token n-grams. The entire system has been built under Sautrela framework [1].

2.1.1. Feature extraction

The system uses Mel-Frequency Cepstral Coefficients (MFCC) as acoustic features, computed in frames of 25 ms at intervals of 10 ms. Neither frame energy nor dynamic features are used.

2.1.2. Tokenization

The tokenization is based on a 200-order GMM [2]. For each 10-dimensional MFCC vector, the tokenizer returns the index of the mixture component yielding the highest score. Three different tokenizers are trained for the submitted tests (Spanish, Mandarin and Hindustani dialect recognition).

2.1.3. N-gram Language Models

Token sequences are modelled using back-off smoothed bigram language models [3]. Language models are initialized through Maximum-Likelihood Estimation and then reestimated using the Maximum Mutual Information (MMI) criterion:

$$F_{\text{MMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(O_r|L_r)^{k_r}}{\sum_{\forall L} p_\lambda(O_r|L)^{k_r} p(L)}$$

where $p_\lambda(O_r|L_r)^{k_r}$ is the likelihood of the acoustic sequence O_r, given the correct language, L_r, and model parameters λ. The exponent $k_r = \frac{C}{length(O_r)}$ mirrors the fact that the language model is not modelling full token sequences, but token *sub-sequences*. The constant C is not heuristically fixed, but chosen by maximizing the Mutual Information function:

$$C = \arg\max \{F_{\text{MMI}}(\lambda)\}$$

To perform MMI training, the rational function growth transformation by Gopalakrishnan [4] is applied. Note that the MMI training criterion attempts to extract as much information as possible from the training data, which includes the underlying language priors [5]. To compensate for the bias implicitly induced by language priors in MMI estimations, model-derived priors are estimated by accumulating posterior probabilities for all the training observations:

$$p_\lambda(L_i) = \frac{1}{R} \sum_{r=1}^{R} p_\lambda(L_i|O_r)^{k_r}$$

Finally, the resulting prior estimates are used to normalize the posterior probabilities:

$$\widetilde{p}_\lambda(L_i|O_r) = \frac{\frac{p_\lambda(L_i|O_r)^{k_r}}{p_\lambda(L_i)}}{\sum_{\forall L} \frac{p_\lambda(L|O_r)^{k_r}}{p_\lambda(L)}}$$

[1]This work has been partially funded by the Basque Government, under program SAIOTEK, projects S-PE05IK06, S-PE05UN32 and S-PE06UN48, and the University of the Basque Country, under project EHU06/96.

2.1.4. Scoring

To decide whether or not the input utterance O_r corresponds to the target language L_T, the normalized posterior probability of the target language $\widetilde{p}_\lambda\left(L_T|O_r\right)$ is compared to the corresponding prior $\widetilde{p}_\lambda\left(L_T\right)$:

$$\text{answer}\left(L_r = L_T\right) = \begin{cases} \text{true} & \text{if} \quad \widetilde{p}_\lambda\left(L_T|O_r\right) \geq p\left(L_T\right) \\ \text{false} & \text{otherwise} \end{cases}$$

which can be rewriten as a likelihood ratio:

$$\text{answer}\left(L_r = L_T\right) = \begin{cases} \text{true} & \text{if} \quad \frac{\widetilde{p}_\lambda(L_T|O_r)}{p(L_T)} \geq 1 \\ \text{false} & \text{otherwise} \end{cases}$$

2.2. Training data

Only the training data provided by NIST/LDC for the LRE07 evaluation were used to estimate the models. All the available data from each test-set were used. All the parameters were automatically estimated from the training data. No heuristics were used.

2.3. Processing speed

Expriments were carried on a dual AMD dual core 270 Opteron server (2.0 Ghz) with 6GB of Ram. As the server can run up to four threads simultaneously and Sautrela framework is based in Java™(the garbage collector runs in parallel), the processing speed was measured runing four simultaneous tests. The memory ussage was also limitted to 300MB per java virtual machine. The resulting runtime factor was 0,0058xRT (the full test of 67 hours was ended in 23 min).

3. References

[1] M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework", in Proceedings of the ASRU Workshop, 2005.

[2] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., "Approaches to language identication using gaussian mixture models and shifted delta cepstral features", in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2002, pp. 89–92.

[3] G. Bordel, A. Varona, and M. I. Torres, "K-TLSS(S) Language models for speech recognition", Proc. ICASSP'97, pp. 819–822, April 1997.

[4] P. S. Gopalakrishnan, D. Kanevski, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems", IEEE Transactions on Information Theory, vol. 37, no. 1, pp. 107-113, January 1991.

[5] Pavel Matejka, Lukas Burget, Petr Schwarz, and Jan Cernocky, "Brno University of Technology System for NIST 2005 Language Recognition Evaluation", 2006 IEEE Odyssey: The Speaker and Language Recognition Workshop, June 2006.