AUTOMATIC LANGUAGE IDENTIFICATION SYSTEM

Jan Černocký, Pavel Matějka, Lukáš Burget and Petr Schwarz

Speech@FIT group, Faculty of Information Technology
Brno University of Technology, Czech Republic
{cernocky,matejkap,burget,schwarzp}@fit.vutbr.cz

ABSTRACT

This paper presents the language identification (LID) system developed in Speech@FIT. The system consists of two parts: Acoustic LID determines the language directly on the basis of features derived from the speech signal. We have improved existing approaches by adding discriminative training of acoustic models. In phonotactic LID, speech is first transcribed by phoneme recognizer into strings or graphs (lattices) of phonemes. On these, "language" models are trained to capture statistics of sequences of phonemes. We have pioneered the use of so called "anti-models" for this task. All experimental results are reported on standard NIST 2003 data; comparison with other published results is favorable to our system.

1. INTRODUCTION

Automatic language identification (LID) has increasing importance among speech processing applications. It can be used to route calls to human operators (commerce, emergency), pre-select suitable speech recognition system (information systems) and has many uses in security applications.

The goal for Language Identification is to determine the language a particular speech segment was spoken. The algorithms for LID can be roughly divided [1] into two groups. In *phonotactic modeling*, a tokenizer transcribes the input speech into phonemes and the scoring is performed on phoneme strings or lattices. This approach is mostly referred to as PRLM (Phoneme recognizer followed by language model) or PPRLM (Parallel PRLM). In *acoustic modeling*, the input features are modeled directly by Gaussian mixture models (GMM), artificial neural networks, support vector machines, or other techniques [2]. This paper deals with both acoustic and phonotactic approaches.

In phonotactic approach [3], the quality of PRLM and PPRLM heavily depends on the quality of phoneme recognition and on the amount of available training data. We use high-quality phoneme recognizer based on so called LC-RC FeatureNet approach and in [3], we have presented phoneme recognizers trained on 4 languages from SpeechDat-East database [4]. Although none of these languages is equivalent to any of the target languages in NIST 2003 LID data, the simple fact that these databases contain 10× more data than OGI-Stories (usually used to train tokenizers in LID) greatly improves the LID accuracy. We extended this work by using phoneme lattices rather than strings for both training and scoring by phonotactic models. This approach was pioneered by LIMSI [6] with good results and our results with phoneme lattices (though our approach was simpler) were also very satisfactory. We further extended PRLM (PPRLM) by using of anti-models [8] — phonotactic models trained on misrecognized segments that should help to discriminate between target and non-target language. Similar approach was used by SRI in large vocabulary continuous speech recognition (LVCSR) [7] to compensate for hypothesis that are acoustically confusable with the correct transcriptions, we have however not seen any use of such technique in LID.

Our acoustic modeling using GMM complements our successful PPRLM [3, 8]. In acoustic modeling, we were inspired by the advantages brought by discriminative training into large vocabulary continuous speech recognition (LVCSR) systems [21].

The paper is organized as follows: section 2 reviews the architecture of our acoustic system, where at first features used for recognition and then acoustic modeling are described. The following section 3 concentrates on the phonotactic system. Section 4 describes fusion and normalization of scores from separate systems. Section 5 presents the data and reports the evaluation and results. The paper is concluded in section 6.

2. ACOUSTIC SYSTEM

2.1. Features

The most widely used features for LID (as well as for other speech processing techniques) are Mel-Frequency Cepstral Coefficients (MFCC). The works of Torres-Carasquillo [9] and others have however shown the importance of broader temporal information for LID. The shifted delta cepstra

This work was partially supported by EC project Augmented Multi-party Interaction (AMI), No. 506811, Grant Agency of Czech Republic under project No. 102/05/0278 and by Ministry of Defense of Czech republic.

(SDC) features are created by stacking delta-cepstra computed across multiple speech frames. The SDC features are specified by a set of 4 parameters: N, d, P and k, where N is the number of cepstral coefficients, d is the advance and delay for the delta-computation, k is the number of blocks whose delta-coefficients are concatenated to form the final feature vector, and P is the time shift between consecutive blocks. In case we denote the original features $o_h(t)$[1], shifted deltas are defined:

$$\Delta o_h(t) = o_h(t + iP + d) - o_h(t + iP - d)$$

for $i = 0, P, 2P, \ldots, (k-1)P$. Obviously, these feature vectors are heavily correlated — most elements are merely copied from one vector to another when we go from t to $t+1$.

Two widely used enhancements of features for LID are RASTA filtering of cepstral trajectories ensuring channel normalization [1] and vocal-tract length normalization (VTLN) [10] which is a simple speaker adaptation.

2.2. Acoustic modeling

Language recognition can be seen as a classification problem with each language representing a class. The most straightforward way to model class s is to construct a Gaussian mixture model that represents feature vectors by a weighted sum of M multivariate Gaussian distributions:

$$p_\lambda(\mathbf{o}(t)|s) = \sum_{m=1}^{M} c_{sm} \mathcal{N}(\mathbf{o}(t); \boldsymbol{\mu}_{sm}, \boldsymbol{\sigma}^2_{sm}) \qquad (1)$$

where $\mathbf{o}(t)$ is the input feature vector and the parameters λ of model of s-th class are c_{sm}, $\boldsymbol{\mu}_{sm}$ and $\boldsymbol{\sigma}^2_{sm}$: mixture weight, mean vector and variance[2] vector respectively. The log likelihood of utterance $\mathcal{O}_r$ given class s is then defined as:

$$\log p_\lambda(\mathcal{O}_r|s) = \sum_{t=1}^{T_r} \log p_\lambda(\mathbf{o}(t)|s) \qquad (2)$$

where T_r is the number of feature vectors in $\mathcal{O}_r$.

In the standard *Maximum Likelihood* (ML) training framework, the objective function to maximize is the total (log) likelihood of training data given their correct transcriptions:

$$\mathcal{F}_{ML}(\lambda) = \sum_{r=1}^{R} \log p(\mathcal{O}_r|s_r) \qquad (3)$$

where λ denotes the set of model parameters, $\mathcal{O}_r$ is r-th training utterance, R is the number of training utterances and s_r is the correct transcription (in our case the correct language identity) of the r-th training utterance. To increase the objective function, the GMM parameters are iteratively estimated using well known Baum-Welch re-estimation formulae (see for example [11]).

In *discriminative training*, the objective function is designed in such a way that it is (or is believed to be) better connected to the recognition performance. One of the most popular discriminative training techniques is Maximum Mutual Information (MMI) training where the objective function is the posterior probability of correct label:

$$\mathcal{F}_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\mathcal{O}_r|s_r)\mathcal{P}(s_r)}{\sum_{\forall s} p_\lambda(\mathcal{O}_r|s)\mathcal{P}(s)}. \qquad (4)$$

We consider the prior probabilities of all classes equal and drop the prior terms $\mathcal{P}(s_r)$ and $\mathcal{P}(s)$. The denominator $\sum_{\forall s} p(\mathcal{O}_r|s)$ is the likelihood of utterance $\mathcal{O}_r$ given the "competing" model representing all possible transcriptions (in our case all language labels). The derivation of parameter update formulae is described in detail for example in [12].

The advantages of MMI are the following:

- It concentrates on precise modeling of decision boundary and does not waste the parameters on highly overlapped features with low discriminative power (Fig. 1).

- It optimizes parameters for good recognition of whole segments (not individual frames) and therefore takes into account the enormous importance of correct speech segmentation. We used segmentation generated by our phoneme recognizer (see Sec. 3.1), where all phonemes are linked to 'speech' class and pause and speaker noises make the 'silence' class. The silence segments are not used for training and testing.

The drawback of MMI is that it also learns the (undesirable) language priors from training data. We equalize the amounts of training data per language but rather than throwing out training data, we appropriately weigh segments in MMI re-estimation formulae.

Discriminative training techniques lead to consistent improvement in accuracy of LVCSR systems [13, 12]. To our knowledge, MMI training of GMMs has not been tested in LID so far. Dan and Bingxi [14] report results with Minimum classification error (MCE) criterion for the training, but the improvement they obtained was less than reported in our paper. We have tested MCE training too, but the results were also not satisfactory.

Our work on MMI training for LID was facilitated by the experience with discriminative training applied in AMI-LVCSR system[3] [15]. We could also rely on our HMM toolkit STK[4] that implements MMI and other discriminative training techniques.

[1] $o_h(t)$ denotes the h-th element of feature vector $\mathbf{o}(t)$

[2] we assume diagonal covariance matrices that can be represented by variances.

[3] AMI is EC-sponsored project Augmented Multi-Party Interaction, http://www.amiproject.org

[4] http://www.fit.vutbr.cz/speech/sw/stk.html

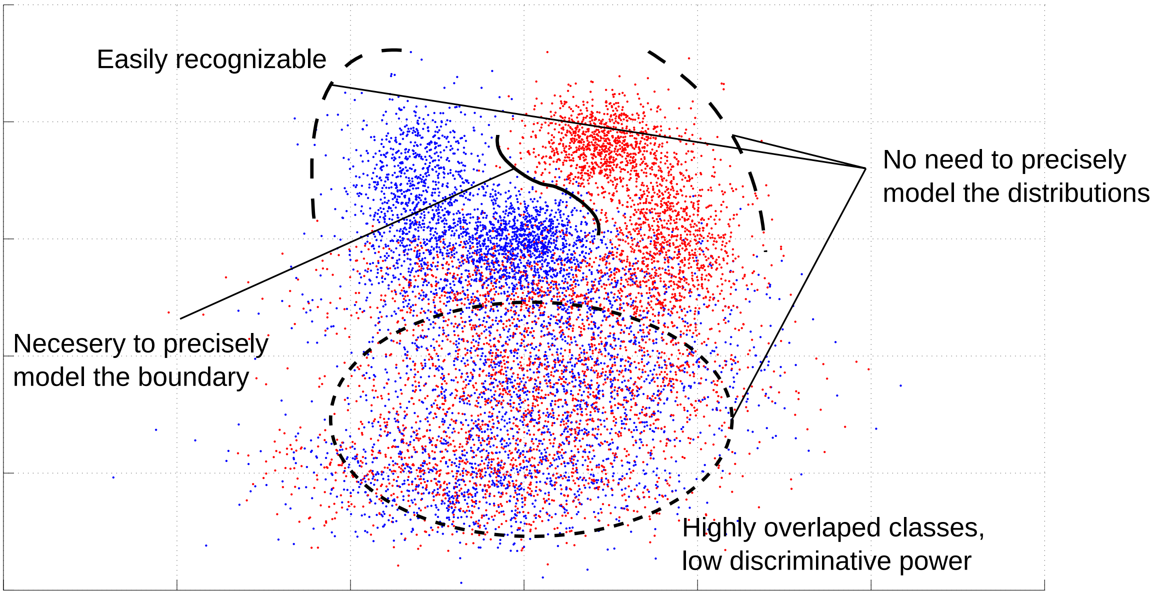

Fig. 1. Highly overlapped feature distribution - differences between ML and MMI training

3. PHONOTACTIC SYSTEM

3.1. Phoneme recognizer

We use a hybrid system based on Neural Networks (NN). The feature extraction makes use of long temporal context. First, Mel filter bank energies are obtained in conventional way. After sentence mean normalization in each band, temporal evolution of critical band spectral densities are taken around each frame. Based on our previous work in phoneme recognition [16, 17], the context of 31 frames (310 ms) around the current frame was selected. This context is split into 2 halves: Left and Right Contexts (hence the name “LCRC”). This allows for more precise modeling of the whole trajectory while limiting the size of the model (number of weights in the NN) and reducing the amount of necessary training data. Both parts are processed by discrete cosine transform to decorrelate and reduce dimensionality. Two NNs are trained to produce phoneme-state posterior probabilities for both context parts. Third NN functions as a merger and produces final set of phoneme-state posterior probabilities (Figure 2). All neural networks [5] have 1500 neurons in hidden layer.

A simple Viterbi decoder from our STK toolkit without any language model constraints processes output of the merger and produces string of phonemes. Phoneme lattices are generated using HTK toolkit[6].

3.2. N-gram language modeling

Smoothed trigram back-off language model was used to capture phonotactic statistics of each language. It was created by passing training speech of all target languages through phoneme recognizer and counting trigrams for each language separately. Phoneme insertion penalty (PIP) in the decoder was tuned on our development set with the best LID performance as criterion. We use standard Witten-Bell discounting [18] implemented in SRI LM toolkit [7] [19].

3.3. Lattices

Since phoneme recognizer is not 100% accurate on 1 best decision, it is advantageous to use richer structure at the end of decoder: lattices instead of strings. At first, acoustic likelihoods contained in lattices are converted to phoneme posteriors. Then, the LM is computed from the new N-gram estimates weighted by these posteriors. Gauvain at al. pioneered this for LID [6].

3.4. Anti-models

Anti-model is a language model modeling the space where target model makes mistakes [8]. Its training works in the following way: we will denote all utterances belonging to language L as set S_L^+ and all utterances not belonging to language L as set S_L^-. First, the training of phonotactic model LM_L^+ of each language L is done in standard way using only

[5] All nets are trained using QuickNet from ICSI http://www.icsi.berkeley.edu/Speech/qn.html

[6] http://htk.eng.cam.ac.uk/

[7] http://www.speech.sri.com/projects/srilm/

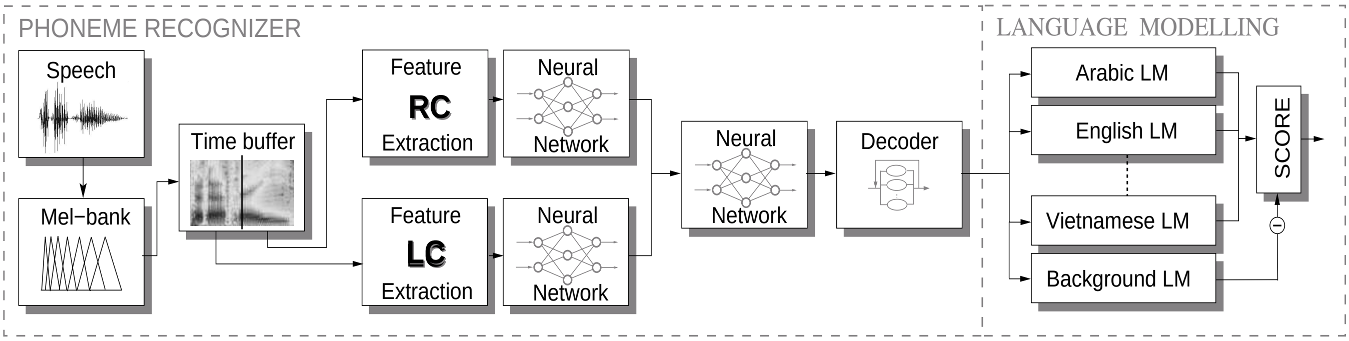

Fig. 2. PRLM system based on phoneme recognizer with split temporal context

the set S_L^+. Then, all *training* utterances are scored by all phonotactic models and posteriors of utterances are derived:

$$\mathcal{P}(\mathcal{O}_r|L) = \frac{\mathcal{L}(\mathcal{O}_r|LM_L^+)}{\sum_{\forall L}\mathcal{L}(\mathcal{O}_r|LM_L^+)} \qquad (5)$$

where $\mathcal{O}_r$ is the r-th training utterance and $\mathcal{L}(\mathcal{O}_r|LM_L^+)$ is the likelihood provided by phonotactic model LM_L^+.

For language L, the parameters of anti-model LM_L^- should be trained on all segments from S_L^- mis-recognized as L. We can however use *all* utterances $\mathcal{O}_r \in S_L^-$ and weight their trigram counts by the posteriors $\mathcal{P}(\mathcal{O}_r|L)$. Obviously, an utterance from S_L^- with high probability to be mis-recognized as L will contribute more to the anti-model than an utterance correctly recognized as language G where $G \neq L$.

Final score of utterance $\mathcal{O}_r$ is obtained by subtracting the weighted likelihood of anti-model from the target model:

$$\log\mathcal{S}(\mathcal{O}_r|L) = \log\mathcal{L}(\mathcal{O}_r|LM_L^+) - k\ \log\mathcal{L}(\mathcal{O}_r|LM_L^-), \quad (6)$$

where the constant k needs to be tuned experimentally.

4. FUSION AND NORMALIZATION OF SCORES

To fuse scores from separate systems, a simple linear combination is done according to:

$$\begin{aligned} score &= \alpha\, GMM_{MMI} + \beta\, PRLM_{HU} + \\ &+ \gamma\, PRLM_{RU} + \delta\, PRLM_{CZ} \end{aligned} \qquad (7)$$

where weights $\alpha, \beta, \gamma, \delta$ are tuned by simplex method on development set.

To obtain the final score of language L, we perform the following normalization using likelihoods of all individual language detectors:

$$\log P(L|\mathcal{O}) \approx \log p(\mathcal{O}|L)/T - \log\sum_l p(\mathcal{O}|l)/T, \qquad (8)$$

where $\log p(\mathcal{O}|L)$ is log-likelihood of speech segment $\mathcal{O}$ given by GMM or LM for language L and T is either number of frames (for GMMs) or phonemes (for PRLM) in speech segment $\mathcal{O}$. The term $\log\sum_l p(\mathcal{O}|l)/T$ can be interpreted as background model.

5. EXPERIMENTS

5.1. Databases

All data used for experiments were recorded over telephone line.

The **phoneme recognizers** used throughout this paper were trained on Hungarian, Russian and Czech SpeechDat-East [4] which performed the best in our previous work [3]. Only phonetically balanced items were used for the training of phoneme recognizers.

Phonotactic language models and **acoustic models** were trained on the CallFriend [5] containing telephone speech of 15 different languages or dialects. Each of 12 target languages (Table 1) contains 20 complete half-hour conversations.

Test Data comes from NIST 2003 LID evaluation [20]. This data set consists of 80 segments with durations of 3, 10 and 30 second in each of 12 target languages (Table 1). All results in this paper are reported for 30s segments. This data comes from conversations collected for the CallFriend Corpus but not included in its publicly released version. In addition, there are four additional sets of 80 segments of each duration selected from other LDC conversational speech sources, namely Russian, Japanese, English and cellular English.

Development Data comes from NIST 1996 LID evaluation and has similar structure.

5.2. Evaluation

The evaluation is done according to NIST [20] per-language, considering each system is a language *detector* rather than recognizer. A standard detection error trade-off (DET) curve is evaluated as a plot of probability of false alarms against the probability of misses with the detection threshold as parameter and equal priors for target and non-target languages.

Arabic (Egyptian)	Japanese	Farsi
French (Canadian French)	German	Hindi
English (American)	Korean	Mandarin
Spanish (Latin American)	Tamil	Vietnamese

Table 1. The twelve target languages in NIST 2003 LID evaluations.

System	EER [%]
PRLM string	3.08
PRLM+lattice	2.25
PRLM+lattice+anti.m.	1.83
PPRLM+lattice+anti.m.	1.42
GMM-ML 2048	4.8
GMM-MMI 128	1.92
Fusion BUT 2006=PPRLM+GMM-MMI	**0.92**
MIT-FUSE	2.8
LIMSI-NN	2.7
BUT-SPDAT 2005	2.4

Table 2. EER of different system for NIST LRE 2003 for 30sec condition

Equal error rate (EER) is the point where these probabilities are equal. The total EER of the whole LID system is the average of language-dependent EERs.

5.3. Results

Table 2 summarizes the results. Conventional phonotactic system using strings of phonemes based on Hungarian phoneme recognizer performs with EER=3.08%. Replacing string output representation by lattices yields an improvement of more than 1% absolute: EER=2.25%. With anti-models, the EER drops to 1.83% (the constant in Eq. 6 was set to $k = 0.3$). Fusing scores from 3 phoneme recognizers brings EER=1.42% which is 23% relative EER reduction.

For GMM model, only segments labeled 'speech' by Czech phoneme recognizer were used. We used 128-component GMM trained under Maximum Mutual Information framework. This system (EER=1.92%) proves its superiority over the state-of-the-art highly dimensional (2048-component) GMM (EER=4.8%) trained under conventional Maximum Likelihood (ML) framework [21].

6. CONCLUSION

Table 2 compares our system to the best results published on NIST 2003 data:

- MIT system [2] labeled MIT-FUSE was based on merging of outputs of PPRLM (6 languages from OGI Stories), Gaussian Mixture Model and Support Vector Machine trained on acoustic features.
- LIMSI-NN system [6] is a PPRLM trained on 3 languages (CallHome – Arabic, SwitchBoard – English and CallHome – Spanish); it uses phoneme lattices to train and score phonotactic models and neural-net based merging of individual scores.
- Our system BUT-SPDAT 2005 [3] is a PPRLM trained on 4 languages from SpeechDat-East with linear merging of individual scores.

The system described in this paper — BUT 2006 — includes PPRLM system based on Hungarian, Russian and Czech phoneme recognizers and acoustic system based on 128-component GMM trained under Maximum Mutual Information framework.

We see that our Hungarian PRLM and Acoustic GMM-MMI as stand-alone systems significantly outperform the other published systems. Merging these systems further improves the performance.

7. REFERENCES

[1] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech.," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[2] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds, "Acoustic,phonetic,and discriminative approaches to automatic language identification," in *Proc. Eurospeech*, Sept. 2003, pp. 1345–1348.

[3] P. Matějka, P. Schwarz, J. Černocký, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," Sept. 2005, pp. 2237–2241.

[4] SpeechDat-E: http://www.fee.vutbr.cz/SPEECHDAT-E.

[5] CallFriend Corpus: http://www.ldc.upenn.edu/Catalog.

[6] J.L. Gauvain, A. Messaoudi, and H Schwenk, "Language recognition using phone lattices," in *Proc. International Conferences on Spoken Language Processing (ICSLP)*, Sept. 2004, pp. 1283–1286.

[7] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V.R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The sri march 2000 hub-5 conversational speech transcription system," in *Proceedings NIST Speech Transcription Workshop*, College Park, MD, USA, May 2000.

[8] P. Matějka, P. Schwarz, B. Burget, and J. Černocký, "Use of anti-models to further improve state-of-the-art prlm language recognition system," in *submitted to Proc. ICASSP*, Toulouse, France, May 2006.

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

[10] J. Cohen, T. Kamm, and A.G. Andreou, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," *J. Acoust. Soc. Am.*, , no. 97, pp. 2346, 1995.

[11] S. Young et al., *The HTK Book*, Cambridge University Engineering Department, 2005.

[12] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, July 2004.

[13] R. Schluter, W. Macherey, B. Muller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," in *Speech Communication*, 2001, vol. 34, pp. 287–310.

[14] Q. Dan and W. Bingxi, "Discriminative training of gmm for language identification," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo,Japan, Apr. 2003, p. MAP8.

[15] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *Proc. NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh,UK, July 2005.

[16] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *submitted to Proc. ICASSP*, Toulouse, France, May 2006.

[17] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Proc. International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, Sept. 2004, pp. 465–472.

[18] I.H. Witten and T.C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Trans. Inform. Theory*, vol. 4, no. 37, pp. 1085–1094, 1991.

[19] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *Proc. International Conferences on Spoken Language Processing (ICSLP)*, Denver, Colorado, Sept. 2002, pp. 901–904.

[20] A.F. Martin and M.A. Przybocki, "NIST 2003 language recognition evaluation," in *Proc. Eurospeech*, Sept. 2003, pp. 1341–1344.

[21] B. Burget, P. Matějka, and J. Černocký, "Discriminative training techniques for acoustic language identification," in *submitted to Proc. ICASSP*, Toulouse, France, May 2006.