IDIAP RESEARCH REPORT

IMPROVING ACOUSTIC BASED KEYWORD SPOTTING USING LVCSR LATTICES

Petr Motlicek Fabio Valente Igor Szoke

Idiap-RR-36-2012

DECEMBER 2012

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

IMPROVING ACOUSTIC BASED KEYWORD SPOTTING USING LVCSR LATTICES

*Petr Motlicek, Fabio Valente**

Idiap Research Institute
Martigny, Switzerland
{motlicek,valente}@idiap.ch

Igor Szoke

Brno University of Technology
Czech Republic
szoke@fit.vutbr.cz

ABSTRACT

This paper investigates detection of English keywords in a conversational scenario using a combination of acoustic and LVCSR based keyword spotting systems. Acoustic KWS systems search predefined words in parameterized spoken data. Corresponding confidences are represented by likelihood ratios given the keyword models and a background model. First, due to the especially high number of false-alarms, the acoustic KWS system is augmented with confidence measures estimated from corresponding LVCSR lattices. Then, various strategies to combine scores estimated by the acoustic and several LVCSR based KWS systems are explored. We show that a linear regression based combination significantly outperforms other (model-based) techniques. Due to that, the relative number of false-alarms of the combined KWS system decreased by more than 50% compared to the acoustic KWS system. Finally, an attention is also paid to the complexities of the KWS systems enabling them to potentially be exploited in real-detection tasks.

Index Terms— KeyWord Spotting (KWS), Spoken Term Detection (STD), Confidence Measure (CM)

1. INTRODUCTION

KeyWord Spotting (KWS) is a technique used to detect keywords (defined a-priori) in speech utterances. Such a technique is essential in spoken document retrieval tasks; the current target users are police and other public/private security authorities.

An acoustic KWS can be seen as a limited vocabulary Automatic Speech Recognition (ASR) system. Unlike ASR, the acoustic KWS does not need to recognize the whole sentence. The keywords are searched in parameterized spoken data (acoustic features) [1]. Unlike acoustic KWS, Large Vocabulary Continuous Speech Recognition (LVCSR) based KWS systems (often called Spoken Term Detection systems) search keywords in the output of the LVCSR, i.e., word recognition strings – lattices [2]. The vocabulary set is usually large, but closed. Therefore, words with low prior probabilities (proper names, etc.) cannot usually be detected in the word recognition lattices and are often denoted as Out-Of-Vocabulary (OOV) words. Possible modifications are provided by hybrid approaches transcribing the speech into lattices of phones or sub-word units which can deal with OOVs. However, the overall KWS detection accuracy is usually lower [3]. Interesting improvements can be achieved by using additional features to boost the confidence scores of the search terms [4], or by combining word and phone indexes [5].

*This work was partially supported by the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management (IM)2"; and by the European commission $7th$ Framework Programme (FP7) ICT Project "Together Anywhere, Together Anytime" (TA2).

The acoustic KWS described above represents a second potential approach towards open-vocabulary KWS, where a model of any keyword (composed of acoustic phone models) is built at the time the keyword is entered [1]. Although this acoustic KWS approach has not been thoroughly investigated and usually performs worse than LVCSR based techniques, its simplicity, speed and robustness to OOVs can be of interest for real-time applications [3].

Due to inaccuracies of ASR technologies used in current KWS, the detected keywords need to be accompanied by Confidence Score (CS) estimates. In the case of acoustic KWS, CSs are estimated as a ratio between likelihood given a keyword model and likelihood given a background model. In the cases of word, sub-word and phone recognition lattices generated by the LVCSR, the confidence can be represented by word, sub-word or phone posterior probabilities, respectively, conditioned on an entire utterance and estimated from those lattices by forward-backward re-estimation [6].

Compared to the acoustic KWS, LVCSR-KWS systems usually yield much better detection performance, however, operating with higher complexity (i.e., far above Real-Time (RT)). This property is particularly valid if an operating point corresponding to a low number of missed keywords is required and there is no constraint on the LVCSR complexity (decoding time) so that rich output recognition lattices can be generated. However, a search in even relatively rich word recognition lattices (generated with low pruning) can lead to observing a small number of False-Alarms (FAs) but an excessively high number of missed keywords. Such the behavior is not acceptable in most current real KWS detection tasks (e.g., security oriented), where a selected operating point usually allows for a higher number of FAs but ensuring a small number of missed words. In these tasks, an acoustic KWS is a more adequate candidate.

In this paper, previously unexplored experimental work on the combination of the acoustic and various LVCSR (word and phone recognition lattice) based KWS systems is reported. We first propose to augment the acoustic KWS by enriching acoustic CS estimates using conventional confidence measures derived from corresponding LVCSR word recognition lattices. Then, several experiments in combining properly aligned individual CSs of various KWS systems using different techniques are conducted. Compared to the acoustic KWS, the best combination yields about 50% decrease in the number of FAs for the pre-defined operating point (ensuring a reasonably low number of missed keywords). During the experiments, we also kept in mind complexity issues of the combined KWS systems due to their potential applicability in a real detection task.

The paper is organized as follows: Section 2 describes the data used in our KWS detection experiments and an evaluation metric used. Section 3 gives more details about employed LVCSR systems, while Section 4 refers to the acoustic KWS. Results on combination of KWS techniques are given in Section 5, followed by discussions.

2. KWS TASK

2.1. Test data

The study is carried out on 16 kHz real unconstrained speech recorded using close-talk microphones in a fairly clean environment (SNR $\sim$ 20dB). In total, about 70 minutes of recordings pronounced in English by non-native (male/female) speakers are used. Due to a chosen scenario, the microphones were not placed in front of each speaker, but rather close to the ears (i.e., to be less visible). This caused some degradation of the recorded speech quality (especially due to a large variation of energy of the speech) and hence renders the KWS task more challenging. In total, 740 occurrences of pre-defined keywords composed of various phone lengths (i.e., 3 to 8 phones) appear in the experimental data and their time positions are precisely annotated.

Due to machine-learning techniques used later in our experiments, a training dataset is required. For this purpose, a subset ($\sim$ 70 minutes) of 16 kHz audio lectures annotated for ASR as well as KWS tasks is employed [7].

2.2. Evaluation metric

Since KWS is a detection task, performance can be characterized by Detection Error Tradeoff (DET) curves of miss (P_{miss}) versus false-alarm (P_{fa}) probabilities. In addition, we also present Equal Error Rates (EERs) – a one number metric often used to optimize a system performance. To highlight achieved detection performances, relative numbers of False-Alarms (FAs) are compared for an arbitrary operating point (which is meaningful for any potential security oriented application). We also present Figure-Of-Merit (FOM) – metric [8], which yields an upper-bound estimate on spoken term detection accuracy averaged over 1 to 10 FAs per hour.

3. LVCSR – KWS

LVCSR used for the KWS detection is a 3-pass AMI[DA][1] system trained on 16 kHz Individual Headset Microphone (IHM) recordings from several meeting corpora (ICSI, NIST, AMI) [9]. In the first pass, PLP features are exploited and Acoustic Models (AMs) represented by HMMs are trained using a Minimum Phone Error (MPE) procedure. In the second pass, Vocal Tract Length Normalization takes place together with Heteroscedastic Linear Discriminant Analysis, MPE and Speaker Adaptive Training (SAT). In the third pass, posterior-based speech features estimated using a Neural Network (NN) system replace PLPs. For the decoding, a 50k dictionary is used together with a 3-gram Language Model (LM). This system reaches a Word Error Rate (WER) of 2.9% on Wall Street Journal (WSJ) Hub2 test set (composed from the November 92-1243 utterances/2.5 hours, 5k dictionary, 3-gram LM).

To compare detection performances of individual KWS systems, the full AMIDA LVCSR system is used so that word recognition lattices are derived in the 3rd pass with "weak" pruning. Overall complexity of the decoding process is about 20xRT. Pre-defined keywords are then searched as an index in the decoded word recognition lattices. Such the resulting KWS system is denoted as KWS^{20xRT}_{LVCSR}. Although the dictionary contains all pre-defined keywords (i.e., no detection of OOVs as searched key-words), the presented KWS scenario does not in fact make difference between OOVs and in-vocabulary words.

[1] http://www.amiproject.org

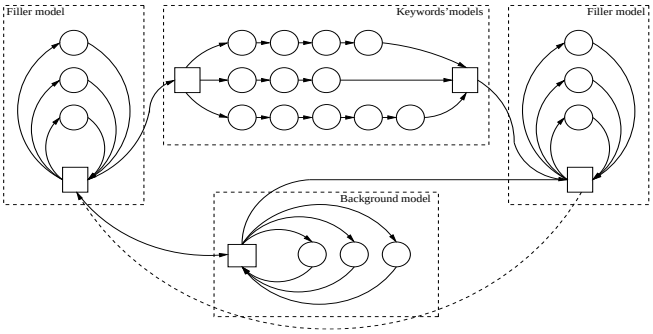

Fig. 1. General scheme of acoustic KWS.

3.1. Simplified systems for fusion

In order to take into account a potential employment of the developed KWS techniques, KWS^{20xRT}_{LVCSR} system based on the 3rd pass and exploiting weak pruning during the decoding is used only to compare final KWS detection performances. In the following experiments, a simplified LVCSR-KWS versions are employed:

(a) 2-pass AMIDA LVCSR system where generation of word recognition lattices is pruned in a way to achieve decoding complexities about 10xRT. Such a system is denoted as KWS^{10xRT}_{LVCSR}.

(b) 2-pass AMIDA LVCSR system generating phone recognition lattices subsequently used in KWS detection. Although the word recognition lattices provide significantly better detection performances than phone lattices (e.g., [10]), the phone lattices show to be useful for the later systems' combination. This system is denoted as $KWS^{10xRT}_{LVCSR_p}$. Similar to the 2nd pass, 3-pass based KWS (with phone lattices) is used only to compare final detection performances (i.e., $KWS^{20xRT}_{LVCSR_p}$ system).

(c) 8 kHz simplified LVCSR system employing the same dictionary and LM as the AMIDA LVCSR. It uses Acoustic Models (AMs) trained in non-discriminative manner without any speaker adaptation technique. AMs are trained on hundreds of hours of Conversational Telephone Speech (CTS) recordings. The KWS detection (denoted as KWS^{5xRT}_{LVCSR}) is done on word lattices and the overall complexity is about 5xRT.

4. ACOUSTIC KWS

As an acoustic KWS, one of currently the best HMM-NN based phone ASR system is employed [11]. More specifically, the phone recognizer exploits context-independent phone models which are represented by phone posteriors estimated using Neural Networks (NNs). For training the NNs, unconventional features (known as TRAPs) are used. TRAPs are derived from relatively long temporal trajectories which are represented by critically band-sized spectral energies. TRAPs are split into two parts – Left and Right Contexts (LC-RC). Outputs of NNs trained separately on LC and RC parts are then merged using another NN called Merger. Merger-NN produces 3-state phone posterior estimates for beginning, center and end of a phone. Such the setting has shown to well estimate overall phone posteriors by precise modeling the whole temporal trajectory while the sizes of NNs are limited. The NNs for generating the phone posterior estimates are trained on a large scale of 16 kHz meeting data.

During KWS detection, word models of searched keywords are created from corresponding phone models (i.e., 3-state phone posterior estimates are transformed into 3-state Hidden Markov Models (HMM) with emission probabilities given by the Merger-NN). Parallelly concatenated keyword models are then accompanied by

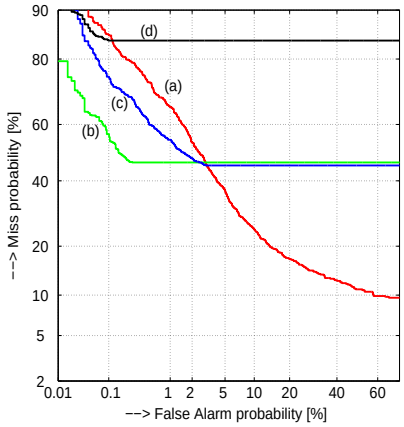

Fig. 2. DET plot – KWS detection performances of different individual systems: (a) KWS^{1xRT}_{acoust}, (b) KWS^{20xRT}_{LVCSR}, (c) KWS^{5xRT}_{LVCSR}, (d) $KWS^{20xRT}_{LVCSR_p}$.

filler and background models (represented by simple phone loops) to create a decoding network, as shown in Fig 1. Likelihoods of the detected keywords are taken from the last state of each keyword model (computed using Viterbi decoder) and compared with the likelihood obtained from the background model. Confidence Score (CS) of each detected keyword is then given as a log-likelihood ratio between these two likelihoods [1]. Such the acoustic KWS is denoted as KWS^{1xRT}_{acoust} and is able to run much faster than LVCSR-KWS (far below 1xRT).

Fig. 2 compares DET curves of the KWS detection on a test dataset using different individual systems. The best versions of LVCSR-KWS systems are shown. The plot indicates that in case of low P_{fa}, KWS^{20xRT}_{LVCSR} significantly outperforms other KWS systems (achieved FOM $\sim$ 37%). However, such the system yields insufficient performances for low P_{miss}. This is caused by the fact that some occurrences of keywords are not found even in weakly pruned word recognition lattices generated in the 3rd pass. Other LVCSR-KWS systems perform worse and yield similar negative properties for low P_{miss}. Acoustic KWS^{1xRT}_{acoust} reaches significantly worse detection performances for low P_{fa}, but can operate up to $P_{miss} \sim 10\%$ (achieved FOM $\sim$ 24%).

4.1. Augmenting acoustic KWS using LVCSR confidence measures

As shown in Fig. 2, the acoustic KWS can operate on much larger scale of DET curve than LVCSR-KWS systems. However, the detection performance is significantly worse, especially due to high number of FAs. In the following experiments, we attempt to improve acoustic KWS by enriching its Confidence Score (CS) estimates using conventional LVCSR confidence measures. Since the LVCSR word recognition lattices are generated for the test dataset, frame-based confidence measures estimated from these lattices can directly be exploited to enrich the acoustic CSs (i.e., of KWS^{1xRT}_{acoust} system).

More specifically, frame-based word entropy $H(W \mid t_n)$ – LVCSR confidence measure – which yields the amount of uncertainty associated with a dictionary W for a given time instance $t = t_n$ is used

$$H(W \mid t_n) = -\sum_i p(W_i \mid t_n)\, log_2\big(p(W_i \mid t_n)\big). \qquad (1)$$

System	EER [%]	FAs [%]
Original **(FOM $\sim$ 24%)**	17.82	100
Enriched **(FOM $\sim$ 25%)**	17.30	91

Table 1. Equal-Error-Rates (EERs) and relative number of False Alarms (FAs) of the original and enriched acoustic KWS^{1xRT}_{acoust} systems for the operating point given by EER ($P_{miss} = 17.82\%$) of the original acoustic KWS.

$p(W_i)$ is a word posterior probability of the hypothesized word W_i (selected from the dictionary) and is computed for each frame from the Acoustic Model (AM) and Language Model (LM) scores of the word recognition lattice using the forward-backward algorithm.

First, for each word w detected by the acoustic KWS and assigned with some acoustic CS (further denoted as $CS_{acoust}(w)$), an LVCSR-CS (denoted as $CS_{LVCSR}(w)$) is computed

$$CS_{LVCSR}(w) = H\big(W \mid t_n \in (t_s, t_e)\big) = \sum_{t_n=t_s}^{t_e} H(W \mid t_n). \qquad (2)$$

t_s and t_e denote starting and end times of each detected keyword w, respectively. Then, $CS_{LVCSR}(w)$ is used as a binary (lenght-independent) threshold for acoustic confidence $CS_{acoust}(w)$ of the given word w

$$CS_{acoust}(w) = \begin{cases} CS_{acoust}(w), & \text{if } CS_{LVCSR}(w) < Thr \\ -\infty, & \text{elsewhere.} \end{cases} \qquad (3)$$

Eq. 3 is valid for log-likelihood ratios used to represent CS_{acoust}. With regards to complexity of the acoustic KWS augmented with CS_{LVCSR}, $H(W \mid t_n)$ is estimated from faster – KWS^{10xRT}_{LVCSR} system, where pruned word recognition lattices are generated in the 2nd pass. Thr is tuned on the training dataset.

Tab. 1 compares detection performances for the acoustic KWS without and with enriching original CS_{acoust} by CS_{LVCSR}. Although EER of the enriched KWS^{1xRT}_{acoust} did not decrease significantly, it yields 9% relative decrease in the number of FAs (computed for the operating point given by EER of the original acoustic KWS).

5. SYSTEM FUSION

In the last experiments, several conventional techniques are exploited to fuse the acoustic and the LVCSR-KWS systems. More specifically, neural network, Maximum Entropy and linear regression techniques are employed to combine Confidence Score (CS) estimates of the keywords detected by hitherto described individual KWS systems. In the first step, CSs are properly aligned so that they correspond to the same keyword w detected in the same time interval $w|_{t_s}^{t_e}$. In the case of non-existing CSs (i.e., a keyword is detected by only some systems), missing CSs are set to $-\infty$, similar to Eq. 3. Then, following techniques for systems' fusion are explored:

NN - A feed-forward backpropagation Neural Network with one hidden layer: A hidden layer comprises 20 nodes with tangent sigmoid as a transfer function. Input is represented by CSs (log based) obtained by individual KWS systems. Output of the NN is trained to discriminate between $0/1$ depending on the true/false occurrence of a given keyword in the transcription. Training of the NN is performed on the training dataset (list of training keywords differs from the list of test keywords).

MaxEnt - Maximum Entropy criterion: It uses conditional maximum entropy models which have been shown to provide good per-

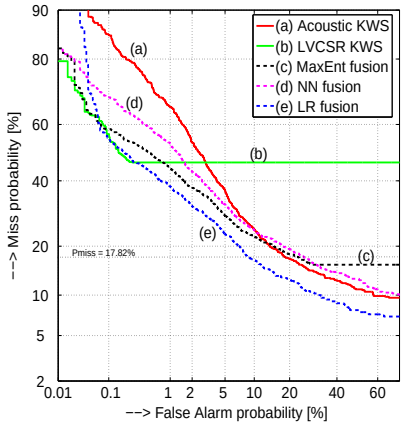

Fig. 3. DET plot – KWS detection performances of combined KWS systems: (a) KWS^{1xRT}_{acoust}, (b) KWS^{20xRT}_{LVCSR}, (c) MaxEnt fusion, (d) NN fusion, (e) LR fusion.

System	EER [%]	FAs [%]
acoustic KWS	17.82	100
MaxEnt	18.54	120
NN	19.32	134
LR	14.46	46

Table 2. Fusion – Equal-Error-Rates (EERs) and relative number of False Alarms (FAs) of combined KWS systems for the operating point given by EER ($P_{miss} = 17.82\%$) of the acoustic KWS.

formance in speech and language processing (language modeling, parsing). Similar to NN, the same training data is employed.

LR - Linear Regression: individual CSs are linearly combined. Resulting CS_{LR} is given as $CS_{LR} = \sum_n a_n \cdot CS_n$, where a_n are weighting constants ($a_n \in (0, 1)$) and CS_n are confidence scores from individual KWS systems. a_n are also estimated on the training dataset. In order to avoid problems with negative infinity values, the LR approach uses posterior probabilities for the combination.

5.1. Experimental results

All NN, MaxEnt and LR classifiers are trained for 4 individual KWS systems: KWS^{1xRT}_{acoust}, KWS^{10xRT}_{LVCSR}, KWS^{5xRT}_{LVCSR} and $KWS^{10xRT}_{LVCSR_p}$. As in the previous experiments, we attempt to employ less complex versions of LVCSR-KWS systems. Achieved DET performances are given in Fig. 3. For the comparison purposes, we also show DET curves of the original acoustic KWS as well as of the best LVCSR-KWS system. EERs and relative number of FAs (computed for the operating point given by EER of the acoustic KWS where $P_{miss} = 17.82\%$) are given in Tab. 2. Combined CSs obtained using NN and MaxEnt classifiers perform better for lower P_{fa} (due to good performances of LVCSR-KWS systems) as shown in Fig. 3. However for lower P_{miss} (as well as for EER-operating point), NN and MaxEnt yield worse performances than simple acoustic KWS. Unlike NN and MaxEnt, the LR classifier significantly improves detection performances over all operating points of the DET curve (achieved FOM $\sim$ 41%).

Tab. 3 shows detection results for the LR fusion. Each individual KWS system increases the overall EER as well as decreases the relative number of FAs. Tab. 3 also compares the LR fusion of the acoustic KWS (and its enriched version) with the LVCSR-KWS systems. The enriched acoustic KWS yields better performances not only as an individual system but also when used in the fusion.

System	EER [%]	FAs [%]
1_a - acoustic KWS **(FOM $\sim$ 24%)**	17.82	100
1_b - enriched acoustic KWS **(FOM $\sim$ 25%)**	17.30	91
1_b + 2	14.59	49
1_b + 2 + 3	14.58	48
1_a + 2 + 3 + 4 **(FOM $\sim$ 40%)**	14.47	49
1_b + 2 + 3 + 4 **(FOM $\sim$ 41%)**	14.46	46

Table 3. LR fusion – Equal-Error-Rates (EERs) and relative number of False Alarms (FAs) of the acoustic and LR fused KWS systems for the operating point given by EER ($P_{miss} = 17.82\%$) of the acoustic KWS: 1_a - original acoustic KWS^{1xRT}_{acoust}, 1_b - enriched acoustic KWS^{1xRT}_{acoust}, 2 - KWS^{10xRT}_{LVCSR}, 3 - KWS^{5xRT}_{LVCSR}, 4 - $KWS^{10xRT}_{LVCSR_p}$.

6. DISCUSSIONS AND CONCLUSIONS

This paper summarizes experimental results achieved with acoustic and LVCSR-KWS systems exploited on conversational audio recordings. The individual LVCSR-KWS systems yield significantly better performances than the acoustic KWS for low number of false-alarms. However for practical applications, an operating point ensuring rather low number of missed keywords is required (e.g., for the security oriented applications). Then, the acoustic KWS outperforms the LVCSR-KWS systems.

Furthermore, keyword confidence scores estimated by the acoustic KWS are enriched by frame-based word entropy – a confidence measure computed from the corresponding LVCSR outputs. Resulting acoustic KWS system is then combined with "relatively" low-complex LVCSR-KWS systems, which yields (in the case of linear regression) a large improvement over any individual system used. Model based combination (NN, MaxEnt) did not bring significant improvements, which was mainly caused by an inequality in the training data. For previously selected operating point, the relative number of false-alarms decreased by more than 50% compared to the acoustic KWS.

7. REFERENCES

[1] I. Szoke, P. Schwarz, L. Burget, M. Karafiat, P. Matejka, J. Cernocky. "Phoneme Based Acoustics Keyword Spotting in Informal Continuous Speech", *in Proc. of TSD 2005*, LNCS/LNAI series, Springer-Verlag, Berlin, September 2005.

[2] J. Garofolo, G. Auzanne, E. Voorhees. "The TREC spoken document retrieval track: A success story". *in Proc. of (TREC-9)*, National Institute of Standards and Technology. NIST, 2000.

[3] I. Szoke, et al. "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", *in Proc. of Interspeech*, pp. 633-636, Lisbon, Portugal, 2005.

[4] C. Parada, A. Sethy, B. Ramabhadran. "Balancing False Alarms and Hits in Spoken Term Detection", *in Proc. of ICASSP*, pp. 5286-5289, Dallas, USA, 2010.

[5] J. Mamou, B. Ramabhadran, O. Siohan. "Vocabulary independent spoken term detection", *in Proceedings of SIGIR*, 2007.

[6] G. Evermann and P. Woodland. "Large Vocabulary Decoding and Confidence Estimation using Word Phoneme Accuracy Posterior Probabilities", *in Proc. of ICASSP*, pp. 2366-2369, Istanbul, Turkey, 2000.

[7] P. Motlicek, P. N. Garner, M. Guillemot and V. Bozzo. "AMIDA/Klewel Mini-Project", Idiap-RR-03-2010, January 2010.

[8] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish. "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting", *in Proc. of ICASSP*, Glasgow, UK, May 1989.

[9] T. Hain, et al, "The AMI System for the Transcription of Speech in Meetings", *in Proc. of ICASSP*, pp. 357-360, Hawaii, USA, 2007.

[10] D. Vergyri et al., "The SRI/OGI 2006 Spoken Term Detection System", *in Proc. of Interspeech*, pp. 2393-2396, Belgium, 2007.

[11] P. Schwarz, P. Matejka, and J. Cernocky. "Towards Lower Error Rates in Phoneme Recognition", *in Proc. of TSD 2004*, LNCS/LNAI series, Springer-Verlag, Berlin, pp. 465-472, September 2004.