IDIAP RESEARCH REPORT

Towards ASR Based on Hierarchical Posterior-Based Keyword Recognition

Petr Fousek [a b]   Hynek Hermansky [a c]

IDIAP–RR 05-64

October 2005

SUBMITTED FOR PUBLICATION

[a] IDIAP Research Institute, Martigny, Switzerland
[b] CTU Prague, Faculty of Electrical Engineering, Prague, Czech Republic
[c] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-64

Towards ASR Based on Hierarchical Posterior-Based Keyword Recognition

Petr Fousek Hynek Hermansky

October 2005

submitted for publication

Abstract. The paper presents an alternative approach to automatic recognition of speech in which each targeted word is classified by a separate binary classifier against all other sounds. No time alignment is done. To build a recognizer for N words, N parallel binary classifiers are applied. The system first estimates uniformly sampled posterior probabilities of phoneme classes, followed by a second step in which a rather long sliding time window is applied to the phoneme posterior estimates and its content is classified by an artificial neural network to yield posterior probability of the keyword. On small vocabulary ASR task, the system still does not reach the performance of the state-of-the-art system but its conceptual simplicity, the ease of adding new target words, and its inherent resistance to out-of-vocabulary sounds may prove significant advantage in many applications.

1 Introduction

Since the early attempts for automatic recognition of speech (ASR), the task has been to recognize words from the closed set of words. As any non-native speaker of the language (or for that matter anybody who may remember the process of acquiring the native language) may testify, human speech communication applying this approach would be impossible. Daily experience suggests that not all words in the conversation, but only a few important ones, need to be accurately recognized for satisfactory speech communication among human beings.

Keyword spotting has a potential to address this issue by focusing only on certain words while ignoring the rest of the acoustic input. Keyword spotting is relatively late discipline in processing of speech and a typical keyword spotting is usually based on a conventional ASR techniques.

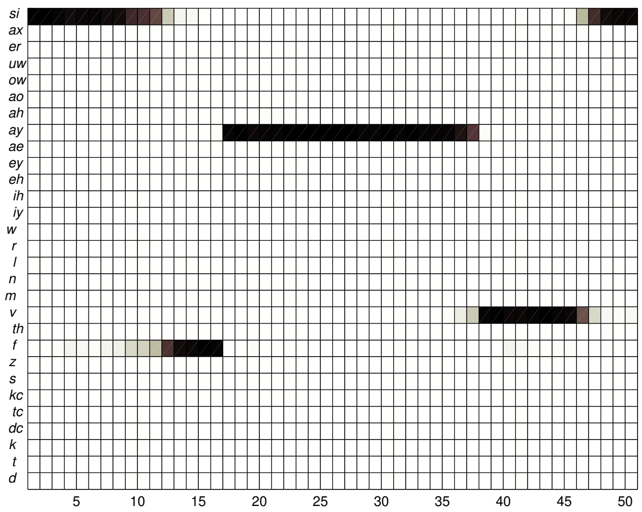

Figure 1: Posteriogram of a word *five* surrounded by silence.

2 The approach

In this work we study an alternative approach, where the ASR task is from its onset seen as a task of recognizing a keyword in a stream of all other sounds while ignoring the rest.

The proposed approach to spotting keywords works in two steps.

1. Equally-spaced posterior probabilities of phoneme classes are estimated from the signal.

2. A probability of a given keyword is estimated from the sequence of phoneme posteriors.

2.1 From speech samples to phoneme posteriors

The first step of the hierarchical processing derives estimates of phoneme posteriors in 10 ms steps from the speech data. This is accomplished as follows: First a critical-band spectral analysis (auditory spectral analysis step from PLP technique [1]) is carried out and a bank of 2-D bandpass filters with varying temporal resolution is applied to the resulting critical-band spectrogram. The resulting 448 dimensional Multi-resolution RASTA (MR) feature vector is fed to a feed-forward artificial neural net classifier, which is trained to give an estimate of posterior probabilities of 29 phoneme classes every 10 ms. An example of time evolution of these posteriors (phoneme posteriogram) for the word ”five” is shown in Fig. 1. More details on MR features can be found in [2].

2.2 From phoneme posteriors to words

Multiple inputs, two-node output multi-layer perceptron neural network (MLP) is used for projecting a relatively long span of the posteriogram (1010 ms) to a posterior probability of a given keyword being present in the center of the time span. Thus, the input to the MLP is a 2929-dimensional vector (29 phoneme posteriors at 100 Hz frame rate). By sliding the 1010 ms window frame-by-frame, the input phoneme posteriogram of an unknown utterance is converted to the keyword posteriogram. A typical keyword posteriogram is shown in the upper part of Fig. 2.

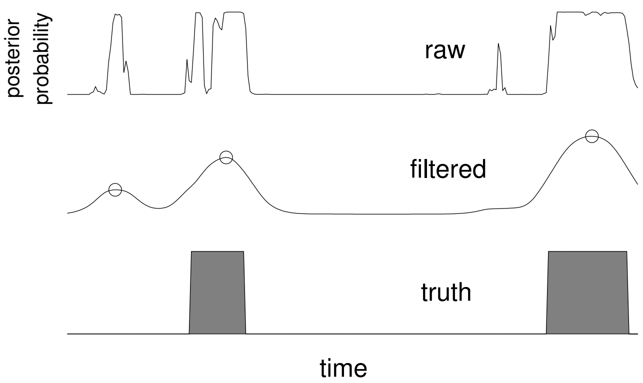

Figure 2: Process of finding position of the keyword.

Even though to human eye the frame-based posterior estimates usually clearly indicate the presence of the underlying word, the step from the frame-based estimates to word-level estimates is very important. It involves nontrivial operation of information rate reduction (carried sub-consciously by human visual perception while studying the posteriogram) where the equally sampled estimates at the 100 Hz sampling rate are to be reduced to non-equally sampled estimates of word probabilities. In the conventional (HMM-based) system, this is accomplished by searching for an appropriate underlying sequence of hidden states.

We have opted for more direct communication-oriented approach where we postulated existence of a matched filters for temporal trajectories of word posteriors, with impulse responses derived by averaging 1 s long segments of trajectories of the respective words, aligned at the word centers. In deriving these averages, we need to deal with cases where the window contains more than one keyword. In the current work, these segments were not included in computing the average. Resulting filters are shown in Fig. 3.

Finally, local maxima (peaks) for each filtered trajectory were found, which indicated that the given word was aligned with the impulse response. The position of the peak indicated the center of the word and a value in the peak was taken as estimate of confidence that the keyword was present.

The process of finding the position of the keyword from its posterior probability is illustrated in Fig. 2 and the whole technique is schematically summarized in Fig. 4.

3 Experiments

The purpose of the experiments was first to test the viability of the proposed approach, second to understand it's properties and tune it's parameters on a development data and third to evaluate and compare the method to ASR system on an unconstrained speech.

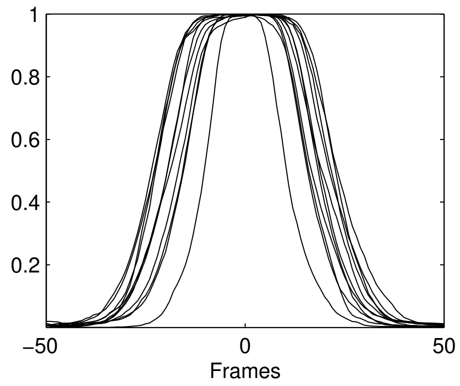

Figure 3: Normalized impulse responses of matched filters for eleven keywords.

3.1 Experiment setup

Two speech corpora were used, OGI-Stories and OGI-Numbers95. Both contain speech recorded over a telephone channel in similar recording conditions. OGI-Stories contains spontaneous continuous speech with rather large vocabulary, OGI-Numbers95 contains strings of digits and numbers [3, 4]. Four distinct data sets were created from these corpora:

training set 1 – 208 files from Stories (2.8 hrs) transcribed on phoneme level by hand,

training set 2 – 2547 files from Numbers95 containing strings of 11 digits from *zero* to *nine* plus *oh* (1.3 hrs) transcribed on phoneme level by hand,

test set – 1433 files from Numbers95 containing strings of 11 digits (1.0 hrs) with word transcription,

extraneous speech set – 129 files from Stories (1.7 hrs) with word transcription.

The neural network estimating phoneme posterior probabilities was trained on merged hand-labeled training sets 1 and 2. We also created binary training targets for keyword spotting networks. In case of training set 2 the aligned word transcription was obtained from automatic alignment using an existing ASR system.

The fact that our test set contains no repeating keywords reveals one particular problem of the system that still needs to be addressed if the system is to be used in certain ASR applications. When two subsequent keywords come, we are likely not to detect both of them. This may not be an issue in some of envisioned application of our system that require merely mark the matching frames containing the keyword but represents problem in ASR evaluations.

3.2 Initial experiment - checking viability

The goal of the initial experiment was to get an idea of limits of the proposed system when applied as a simple speech recognizer. The experiment was evaluated on development set in terms of word error rate (WER), the task was to recognize eleven digits in tested utterances. The baseline was an HMM-based ASR system with context independent phoneme models trained on train set 2. Performance of baseline system was 5.2% WER in case of standard 3×13 PLP features and 3.4% WER in case of phoneme posterior features derived from MR features. Both systems were tuned to give the same number of insertions and deletions.

Eleven independent MLPs, each of which with 1 s long trajectory of phoneme posteriors at the input (2929 features) and two complementary outputs were trained on train set 2 to give frame-wise keyword posteriors for eleven keywords. Next the matched filters for these keyword posterior

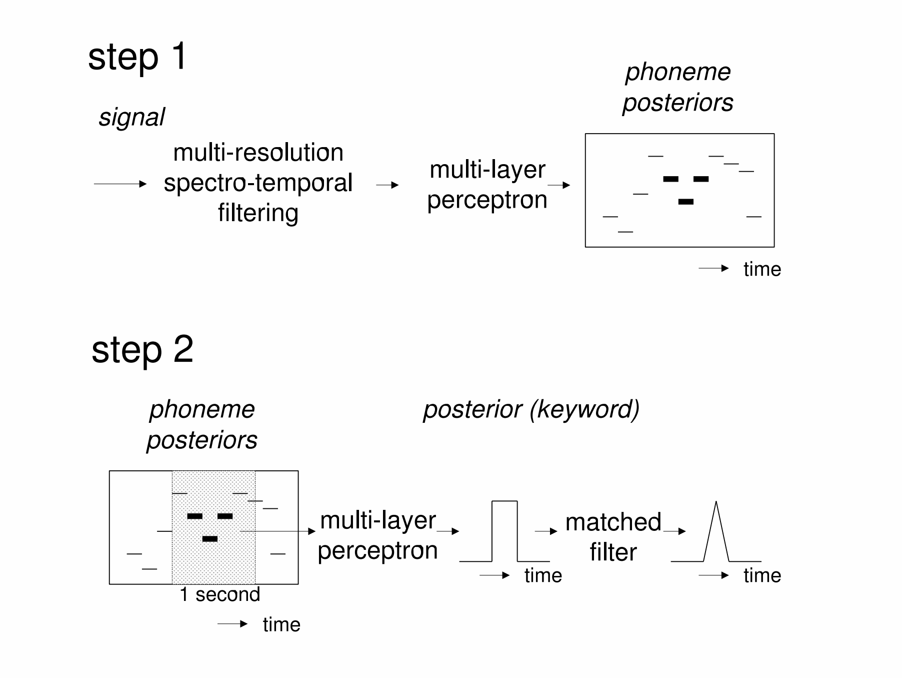

Figure 4: Schematic diagram of the proposed technique.

trajectories were derived by computing the mean trajectory patterns for each of the keywords in the training data. For the first insight we found a fixed threshold on the peak value, which balanced insertions and deletions and yielded encouraging 9.9% WER.

We were also curious whether the system would still work if we omitted the intermediate step of phoneme posteriors and estimated keyword posteriors directly from 448 MR features. The performance dropped to 16.2% WER. Furthermore, when we omitted also the MR feature computation and trained the keyword networks on 1 s trajectory of critical band energies, we got about 31% WER. This suggests that the hierarchical processing employing high-dimensional features and intermediate phoneme classes seem to be beneficial.

The aim of the next experiment was to study the WER with respect to false alarm (FA) rate and to optimize settings for further use. The task was again to recognize 11 digits. For each keyword we found a threshold for alarm so that we get a constant FA rate. Once we found all 11 thresholds for the given FA rate, we evaluated WER for such setting (see the second column of Tab. 1).

system	initial	enhanced	HMM
FA/h	WER[%]	WER[%]	WER[%]
20	19	16	53
25	15	11.6	40
30	12.1	10.3	26
40	9.9	9.3	3.4

Table 1: Word error rate as a function of false alarms per hour of the proposed system on digits recognition task.

We observe that the proposed approach can act as ASR at 12% WER level while keeping at most

30 false alarms per hour. On the other hand, the competitive HMM-based ASR system with its best performance of 3.4% WER yields 38 false alarms per hour and when modified to yield the 30 alarms per hour by manipulating its insertion penalty, its performance degrades to 26% WER! Further efforts for lowering the FA rate in HMM system degraded the performance yet further (see the fourth column in Tab. 1).

3.3 More discriminative training

When training a new set (the improved set) of 11 keyword spotting MLPs on a joint set of train set 2 (Numbers) with a subset of train set 1 (Stories) in a ratio about 1:1, we lowered roughly twice the prior probabilities of all keywords. After re-setting the thresholds we lowered the FA rate while simultaneously improved the WER (see the third column of Tab. 1). This supports the obvious: a sufficient amount of negative examples is necessary in discriminative training of classifiers.

3.4 Unconstrained speech

The task we have been most interested from the onset of our work is the performance in the situation when test data contain a lot of out-of-vocabulary speech. For this we have appended the digits test set by 1.7 hours of extraneous speech set (Stories). Standard HTK-based evaluation procedure was applied as in the previous ASR tasks but the extraneous speech from OGI Stories was labelled as no speech. Results from both the HMM-based system and the enhanced system are shown in Table 2. Number of correctly recognized items follows the trend from the previous experiments. However, there is a huge difference in a number of inserted words. While the HMM system inserts almost 12 000 extraneous words, thus bringing its final WER rather unacceptable 152%, our new system based on parallel key-word spotters inserts only about 1300 words, thus degrading in performance only to 24% WER. Exact numbers are given in Tab. 2.

system	False alarms	WER [%]
ASR baseline	11925	152
enhanced system	1313	24

Table 2: Results on a joint set of Numbers95 digits and unconstrained speech from OGI-Stories.

4 Discussion and conclusions

Our digit-recognizing HMM system represents a typical closed-set vocabulary system, which, when presented with out-of-vocabulary word, attempts to match it with the word from its closed-set vocabulary, yielding a false alarm. This problem is typically addressed by introducing some measure that provides an estimate of confidence in decision about the identity of the underlying word (a difficult research problem on its own).

In this paper we attempted to develop an alternative system based on a set of parallel discriminative classifiers that is better capable of yielding no output when presented with an unknown out-of-vocabulary word and demonstrated that this approach can inherently reduce the insertion error problem on out-of-vocabulary words very significantly.

This has been achieved by a new approach that differs from the current ASR strategies in several aspects:

1. The recognizer for N words is built as a system of N parallel discriminative binary classifiers, each classifying the keyword against the rest of other possible sounds.

2. The classification is based on hierarchical processing where first equally-spaced posterior probabilities of phoneme classes are derived from the signal, followed by estimation of the probability of the given keyword from the sequence of phoneme posteriors.

3. Unlike as in the most current ASR systems, no explicit time warping is done. Instead the binary classifier is trained for word length invariance on many examples of the keyword.

The issue of time warping may deserve some more discussion. It is acknowledged that the introduction of dynamic time warping for the alignment of words with identical phonetic value but of non-equal length was and remains one of important ASR advances. At the same time, however, it may rather arbitrary modify speech dynamics and as such may also be one of essential limiting factors in the current ASR. The system presented in this paper does without this engineering trick and yet seems to be capable to deal at least to some extent with the time issue. This may encourage some more work towards approaches that respect all-important signal dynamics.

5 Acknowledgments

The work was supported by DARPA under the EARS Novel Approaches grant no. MDA972-02-1-0024. The other sources of support were the IM2, AMI and M4 grants.

References

[1] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am., Vol. 87, No. 4, April 1990.

[2] Hermansky, H., P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR", Proc. of Interspeech 2005, Lisbon, Portugal, September 2005.

[3] Cole, R. et al., "Telephone Speech Corpus Development at CSLU", In Proc. of ISCLP '94, pp. 1815–1818, Yokohama, Japan, 1994.

[4] Cole, R. A., M. Noel, T. Lander, T. Durham, "New Telephone Speech Corpora at CSLU", In Proc. of Eurospeech '95, pp. 821–824, Madrid, Spain, 1995.