The Role of Speech in Multimodal Human-Computer Interaction (Towards Reliable Rejection of Non-Keyword Input)

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

IDIAP–RR 05-63

October 2005

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

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 05-63

The Role of Speech in Multimodal Human-Computer Interaction (Towards Reliable Rejection of Non-Keyword Input)

Hynek Hermansky Petr Fousek Mikko Lehtonen

October 2005

Abstract. Natural audio-visual interface between human user and machine requires understanding of user's audio-visual commands. This does not necessarily require full speech and image recognition. It does require, just as the interaction with any working animal does, that the machine is capable of reacting to certain particular sounds and/or gestures while ignoring the rest. Towards this end, we are working on sound identification and classification approaches that would ignore most of the acoustic input and react only to a particular sound (keyword).

1 Introduction

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. The important key-words are more likely to be rare-occurring high-information-valued words. Human listeners can identify such words in the conversation and possibly devote extra effort to their decoding. On the other hand, in a typical automatic speech recognition (ASR) system, acoustics of frequent words are likely to be better estimated in the training phase and language model is also likely to substitute rare words by frequent ones. As a consequence, important rare words are less likely to be well recognized. Keyword spotting by-passes this problem by attempting to find and recognize only certain words in the utterance while ignoring the rest. Doing this in a confident way would open new possibilities in human-computer interaction.

2 Proposed Approach

Since keyword spotting is relatively younger than ASR, it is not clear if the LVCSR-based keyword spotting approaches are a consequence of a simple inertia in engineering where any new problem is seen in the terms of the old one, or the optimal strategy. In this work we study an alternative approach where the goal is to find the target sounds from an acoustic stream while ignoring the rest.

Towards this goal, we propose 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.

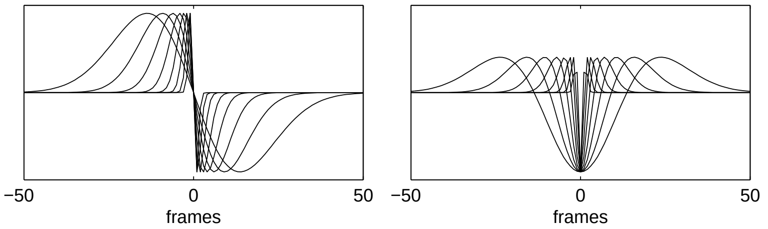

Figure 1: Normalized impulse responses of the two sampled and truncated Gaussian derivatives for σ = 8 − 130 ms.

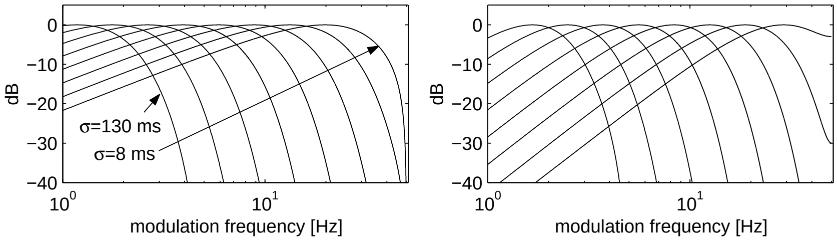

Figure 2: Normalized frequency responses of first two sampled and truncated Gaussian derivatives for σ = 8 − 130 ms.

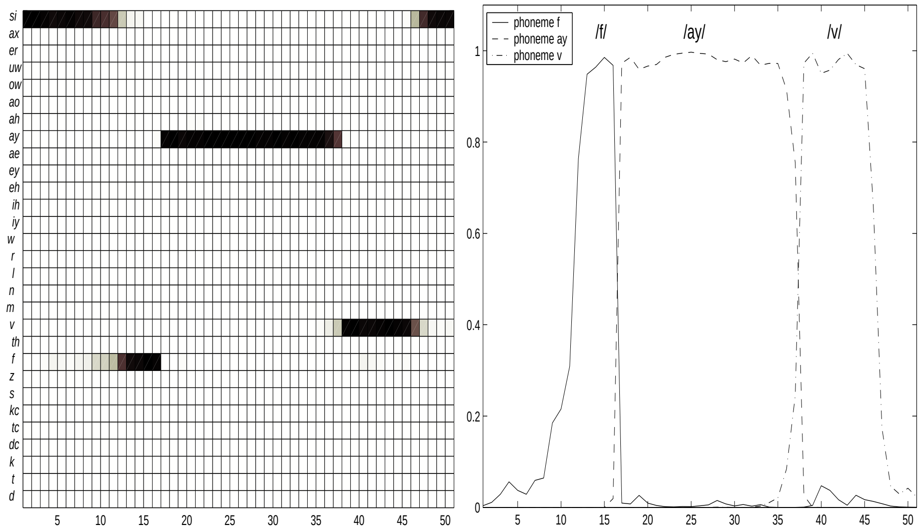

Figure 3: Left: Posteriogram of the word *five* followed by silence. Right: Trajectories of phoneme probability estimates.

3 Steps of Hierarchical Processing

3.1 From Acoustic Stream to Phoneme Posteriors

The first step derives estimates of phoneme posteriors at 10 ms steps (100 Hz sampling rate) 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 and spectral resolutions is applied to the resulting critical-band spectrogram. We implemented the 2-D filtering by first processing a trajectory of each critical band with temporal filters and subsequently applied frequency filters to the result. By filtering temporal trajectory of each critical band with a bank of fixed length low-pass FIR filters representing Gaussian functions of several different widths (determined by standard deviation σ) and by subsequently computing first and second differentials of the smoothed trajectories we would obtain a set of modified spectra at every time step. The same filter-bank is applied to all bands.

In the implementation, we use directly the discrete versions of the first and second analytic derivatives of a Gaussian function as impulse responses. Filters with low σ values yield finer temporal resolution, high σ filters cover wider temporal context and yield smoother trajectories. All temporal filters are zero-phase FIR filters, i.e. they are centered around the frame being processed. Length of all filters is fixed at 101 frames, corresponding to roughly 1000 ms of signal, thus introducing a processing delay of 500 ms. First and second derivatives of Gaussian function have zero-mean by the definition. By using such impulse responses we gain an implicit mean normalization of the features within a temporal region proportional to the σ value, which infers robustness to linear distortions. Since we use discrete impulse responses with a length fixed at 101 samples, we approximate real Gaussian derivatives with certain error, which increases towards both extremes of σ, thus limiting the possible σ values to a range of approximately 6–130 ms. In our experiments we use eight logarithmically spaced impulse responses in a σ range 8–130 ms. These responses representing the first and the second Gaussian derivatives are shown in the left and right parts of Fig. 1, respectively. Related frequency responses are illustrated in Fig. 2.

Temporal filtering of 15 critical band trajectories with a bank of 2×8 filters (two derivatives of the Gaussian at eight different standard deviations) results in 16 modified auditory spectra at every 10 ms step, containing overall $15 \times 2 \times 8 = 240$ features.

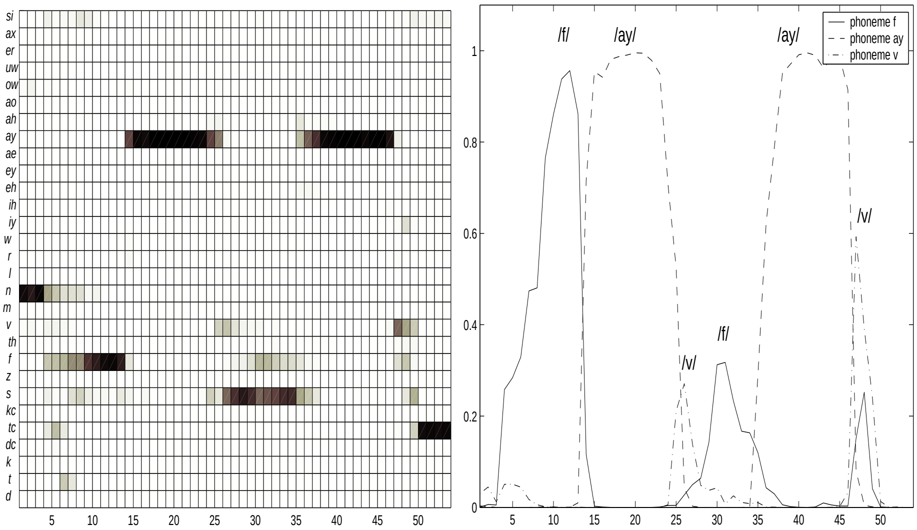

Figure 4: Left: Posteriogram of a sequence *five five* with classification errors and weak posteriors. Right: Trajectories of phoneme probability estimates.

Subsequently we implement the full 2-D filtering by applying frequency filters to the modified auditory spectra. The first frequency derivative is approximated by a 3-tap FIR filter with impulse response {-1.0; 0.0; 1.0}, introducing three-bands frequency context. This time-invariant filter is applied across frequency to each of the 16 modified auditory spectra. Since derivatives for the first and last critical bands are not defined, we obtain (15 - 2)×16 = 208 features.

Final feature vector is formed by concatenating the 240 features from temporal filtering with the 208 features from the full 2-D filtering, thus yielding 448 features. This feature vector is fed to an MLP neural net classifier (TANDEM probability estimator [2]), which is trained to give an estimate of phoneme posterior probabilities at every 10 ms step (phoneme posteriogram). An example of such phoneme posteriogram for the word "five" is shown in Fig. 3. The keyword is in this case easy to spot because the phoneme segments are well classified in the posteriogram. However, this is not always the case. Fig. 4 illustrates a more difficult case for the same word, where the speech rate is higher and there are classification errors. More details of the technique can be found in [3].

3.2 From Phoneme Posteriors to Words

Multiple input, two-node output MLP is used for mapping of relatively long (1010 ms) span of the posteriogram to a posterior probability of a given key-word being within this time span.

Thus, the input to the MLP is a 2929-dimensional vector (29 phoneme posteriors at 100 Hz frame rate). The MLP is trained on the training part of the OGI Digits database (about 1.3 hours of speech), containing 11 digits from zero to nine (including "oh").

In the operation, the input phoneme posteriogram of the unknown utterance is converted to the key-word posteriogram by sliding the 1010 ms window frame-by-frame over the phoneme posteriogram. A typical keyword posteriogram is shown in Fig. 5.

Even though (as illustrated in the figure) 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.

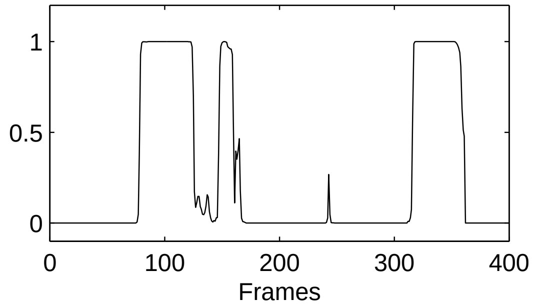

Figure 5: An example of key-word posteriogram.

We have opted for more direct communication-oriented approach where we postulated existence of a matched filter for temporal trajectories of word posteriors, with impulse response 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 key-word. In the current work, these segments were not included in computing the average.

Resulting filter is shown in Fig. 6, and an example of the filtered posteriogram is illustrated in Fig. 7.

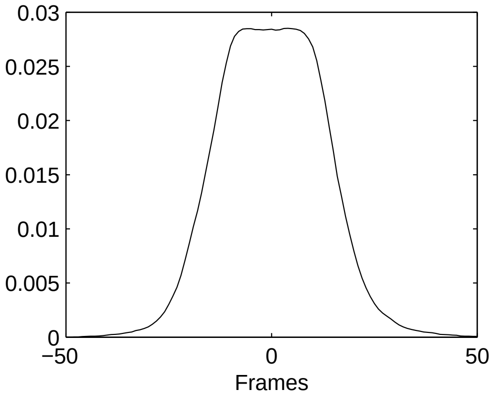

Figure 6: Impulse response of the key-word matched filter.

In the next step, local maxima (peaks) for each filtered trajectory were found. The values in the peaks were taken as estimates of probability that the center of the given word is aligned with the center of the impulse response of the respective matched filter and retained, all other data were discarded. An example of such a peak can be seen in Fig. 7.

The whole technique is schematically illustrated in Fig. 8.

4 Results

As a test we have processed about 104 minutes of speech data containing fluent pronunciation of digit strings (OGI Numbers [4]). Among 12389 digits, there were 1532 ”one”s. 1339 (87.4%) of these were correctly identified, and there were 19 (1.2%) false alarms and 193 (12.6%) misses. Most of misses are caused by smearing short drops in probability due to filtering (as discussed below) that may indicate a succession of several target words. Not counting these errors, the overall detection rate would be 93.4%.

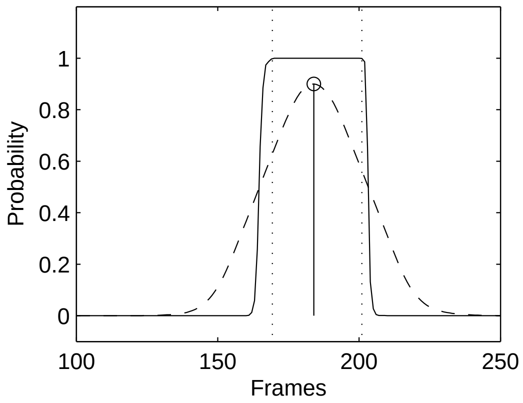

Figure 7: Raw time trajectory of key-word posterior probability (solid line), filtered trajectory (dashed line) and estimated location of the key-word (circle). Region between dotted lines represents a labelled location of the key-word.

5 Summary and Discussion

Hierarchical classification of relatively long chunks of speech signal has been applied to spotting a target word in speech stream. This hierarchical procedure estimates posterior probabilities of phonemes from multi-resolution speech representation obtained by time-frequency filtering of critical-band spectral energy, followed by the estimation of probability of occurrence of the targeted word in the neighborhood of the current time instant. A simple deletion of the speech signal for the instants where the trajectory is below certain threshold allows for retention of most of the targeted speech.

In order to uniquely identify the presence of the underlying target word, a simple procedure is introduced in which the time trajectory of key-word probability is filtered by a box-car shaped matched filter, representing mean length of the key-word and above-threshold peak of this trajectory indicate occurrence of the key-word. The filtering eliminates most of spurious short peaks, thus allowing for simple counting the identified key-words but it also often eliminates short drops in the probability due to several successive key-words which are then counted only as one word.

Result reported in this paper is very preliminary and should be taken only as an indication of feasibility of the proposed approach. Similar approach can be also applied to spotting of phonemes [5] and then digits could be represented as sequences of phonemes. Currently, such an approach yields yet significantly higher word-spotting detection rates.

The presented technique is very straightforward, involves neither any time-warping nor any searches, and can be implemented on-line with a relatively short algorithmic delay. It offers simple but interesting alternative to most of the current speech recognition approaches with a potential for further evolution.

References

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

[2] Hermansky, H., D.P.W.E. Ellis, S. Sharma, “Connectionist Feature Extraction for Conventional HMM Systems”, Proc. of ICASSP 00, Istanbul, Turkey, 2000.

[3] Hermansky, H., P. Fousek, “Multiresolution RASTA filtering for TANDEM-based ASR”, Proc. of Interspeech 2005, Lisbon, Portugal, September 2005.

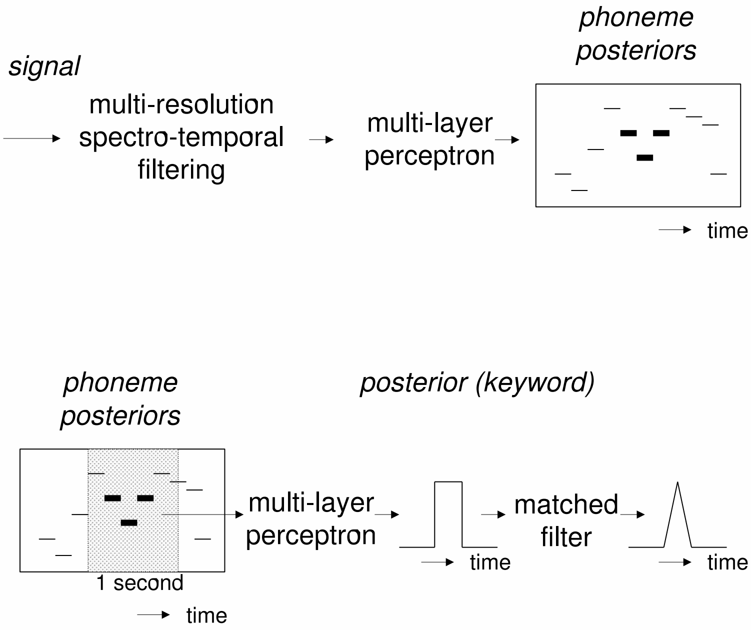

Figure 8: Schematic diagram of the proposed technique.

[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.

[5] Lehtonen, M., P. Fousek, H. Hermansky, “Hierarchical Approach for Spotting Keywords”, IDIAP Research Report, 2005.