Hierarchical Approach For Spotting Keywords

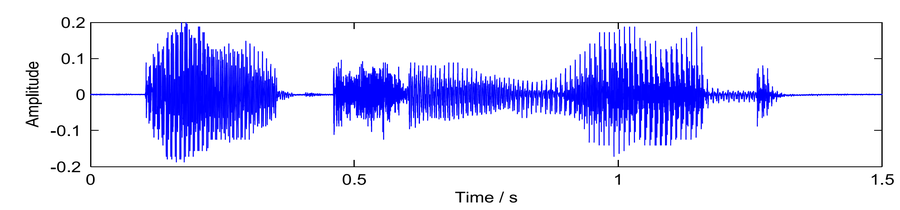
Mikko Lehtonen[1,2] , Petr Fousek[1,2] and Hynek Hermansky[1]

[1] IDIAP Research Institute, Martigny, Switzerland
[2] Helsinki University of Technology, Helsinki, Finland
[3] Czech Technical University in Prague, Prague, Czech Republic

INTRODUCTION

Daily experience suggests that not all words in the conversation, but only a few important ones, need to be recognized for satisfactory speech communication. Keyword spotting approaches this by trying to recognize only a limited number of words while ignoring the rest.

Typical keyword spotting systems are still based on conventional automatic speech recognition (ASR), which might not be the optimal strategy. In this work we study an alternative approach to keyword spotting where the goal is to find the target sounds and ignore the rest.



FROM ACOUSTIC STREAM TO PHONEME POSTERIORS

Phoneme posterior estimates derived every 10ms

Feature extraction using **2-D filtering** of **critical band spectrogram:**

- Temporal filtering with Gaussian filters (Fig. 1)
- Derivatives across frequency

Features are fed to an MLP (TANDEM probability estimator [1]) trained to give the phoneme posterior estimates

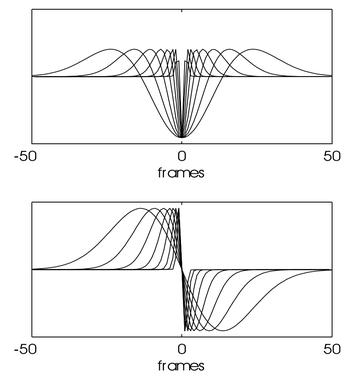


Fig. 1. Normalized impulse responses of the two sampled and truncated Gaussian derivatives.

FROM FRAME-BASED PHONEME POSTERIORS TO PHONEME-SPACED POSTERIORS

Phonemes are found by filtering the posteriogram with a bank of matched filters (Fig. 2)

Matched filters are obtained by averaging 0.5 s long segments of phoneme trajectories

Local maxima of the filtered posteriogram are extracted

A posterior **threshold** is applied

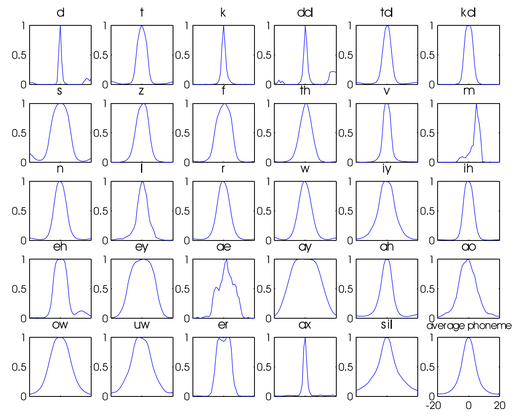


Fig. 2. The matched filter bank of temporal trajectories of phoneme posteriors

FROM PHONEME-SPACED POSTERIORS TO WORDS

An alarm is set, if the right stream of phonemes appears within certain intervals

These intervals are defined by looking at the keywords in the training data (Fig. 3)

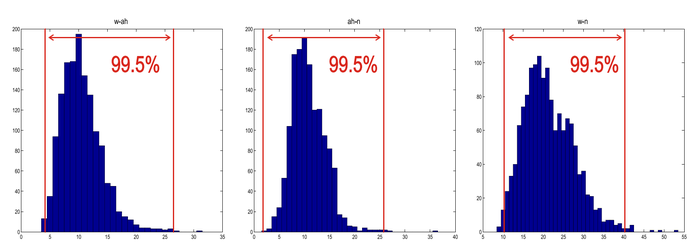


Fig. 3. Histograms of the distances (in frames) between phonemes of the word *one*.

CONCLUSIONS

Our technique looks only for the target sounds and ignores the rest

The approach was applied quite successfully for spotting a digit in the stream of other digits

The performance depends mostly on the accurate frame-level phoneme posterior estimates

In the case of unrestricted speech, the step from phoneme estimates to words needs further development

Keyword	Detection rate 5 FAs/h	Detection rate 10 FAs/h
One	93.4%	95.2%
Two	85.0%	90.9%
Three	95.0%	96.1%
Four	90.4%	92.5%
Five	82.9%	90.5%
Six	94.3%	95.2%
Seven	91.0%	91.6%
Eight	41.6%	51.6%
Nine	46.7%	69.1%
Zero	90.6%	92.5%

Two telephone corpora were used: OGI-Stories [2] and OGI-Numbers95 [3]. The test data was a subset of the latter, containing only digits.

ACKNOWLEDGEMENTS

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811,publication).

REFERENCES

1. Hermansky, H. Et al: Connectionist Feature Extraction for Conventional HMM Systems. In Proceedings of ICASSP'00, Istanbul, Turkey, 2000.
2. Cole R. et al.: Telephone Speech Corpus Development at CSLU. In Proceedings of ISCLP '94, pp. 1815--1818, Yokohama, Japan, 1994.
3. Cole R. et al.: New Telephone Speech Corpora at CSLU. In Proceedings of Eurospeech '95, pp. 821--824, Madrid, Spain, 1995.