

Final Report:

Acoustic Processing Group-WS97

Group Members:

Andreas G. Andreou (CLSP-JHU, USA)

Hynek Hermansky (OGI, USA)

Christian J. Wellekens (EURECOM, France)

Yasuhiro Minami (NTT, Japan)

Terri Kamm (DOD, USA)

Juergen Luetin (IDIAP, Switzerland)

Dan Fain (Caltech, USA)

Sarel van Vuren (OGI, USA)

Group Leader:

Andreas G. Andreou

Electrical and Computer Engineering and
Center for Language and Speech Processing
Johns Hopkins University

3400 N. Charles Street, Baltimore, MD 21218 USA

Telephone: +1-410-516-8361

Fax: +1-410-516-8313

E-mail: andreou@jhu.edu

Abstract

We revisit the architecture of the acoustic processor employed in the state of the art, large vocabulary, continuous speech recognition systems. We investigate data driven processing exploring different techniques at different context scales. At the short time scales (context $\approx 10\text{ms}$) we investigate the non-linear frequency mapping known as Mel-scale. At the medium time scales (context $\approx 100\text{ms}$) we investigate linear discriminant and heteroscedastic discriminant transforms. At time scales with context $\approx 1000\text{ms}$, we investigate novel methods for feature trajectory filtering. At the utterance time scales, ($\approx 500\text{ ms}$ to 4s) we experiment with adaptive, cepstrum bias normalization techniques.

The Mel-scale investigation has established that the natural scale for speech communication (Mel or Bark scale) can be learned from the data using a speech decoding paradigm. Our research with optimal feature extraction has yielded, a global optimal linear transform that can substitute current techniques for incorporating context information and reducing the dimensions of the feature vector. This transform trained on SWITCHBOARD has yielded a 0.8% word error reduction. The work with trajectory filtering and cepstrum bias normalization were encouraging but time limitations did not permit a complete investigation of the different ideas.

1 Introduction

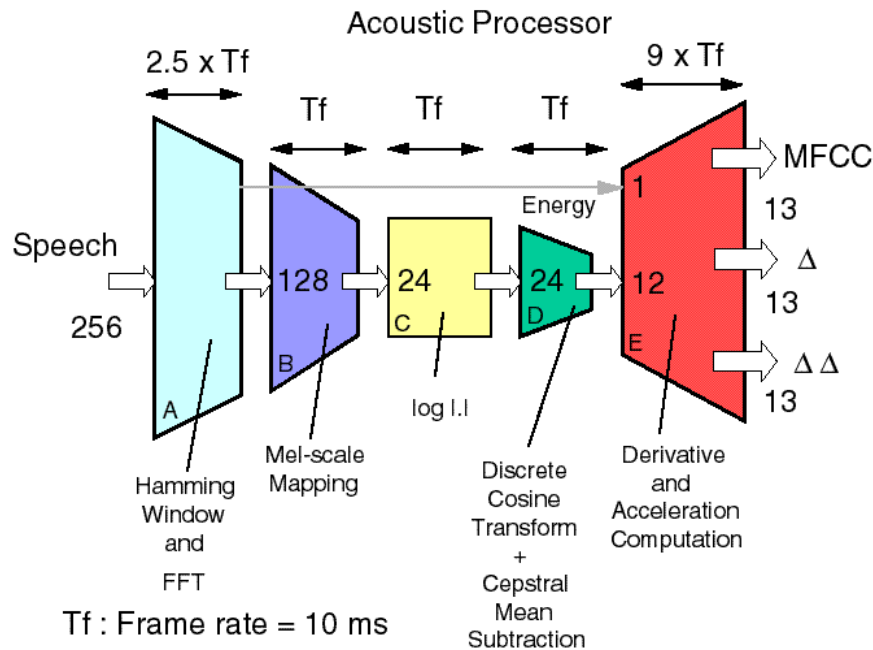


Figure 1: Acoustic processing for automatic speech recognition

Feature extraction in state-of-the-art speech recognition systems, when considered from a data dimensionality viewpoint, is a two step process (see Figure 1. In the first step, the dimensionality of the speech waveform is reduced using cepstral analysis or some other analysis procedure motivated by human speech perception. In the second step, the dimensionality of the obtained features vector is increased by the formation of an extended features vector that includes derivative and acceleration information.

Our group has worked on different aspects of acoustic processing seeking an answer to the following question:

Can we learn from the data how to extract features?

As it will be shown in the remainder of this report the answer to this question is and exciting and affirmative yes.

All at all we have worked as a group in the following sub-projects:

1. Learning the Mel-scale and optimal VTN mapping: *T. Kamm, A.G. Andreou and H. Hermansky*
2. Unsupervised learning of cepstrum-bias on per utterance basis: *Y. Minami*
3. LDA derived feature trajectory filtering and RASTA HMMs *H. Hermansky, J. Luettin and S. van Vuuren*
4. Joint learning of HMMs and feature trajectory filters: *C. Wellekens, H. Hermansky and T. Kamm*
5. Heteroscedastic discriminant analysis: optimal feature filtering: *D. Fain, A.G. Andreou and T. Kamm*

We summarize our work in the following sections and provide more extensive write-ups as Appendices.

2 Learning the Mel-scale and Optimal VTN mapping

This sub-project was a continuation of the work on VTN (Vocal Tract Normalization) from the 1994 Frontiers for Speech Progressing workshop at CAIP. We designed and conducted a set of experiments aimed at learning from data the most discriminative distribution of frequency bins in the acoustic processor. Our experimental results indeed show that the psychophysically motivated Mel-scale is a member of a class of discriminative partitions. For more details please see Appendix 1.

3 Unsupervised learning of cepstrum-bias on per utterance basis

In this sub-project, we seek to improve on the traditional technique of CMN (Cepstral Mean Normalization) by computing the normalization bias on per utterance basis. As it can be seen in Table 1 below, the experiments conducted on a small test set were encouraging but the same degree of success was not obtained on the large set. More details of this work can be found in Appendix 2.

Table 1: Cepstrum bias adaptation

Method Used	Short Test (WER)	Long Test (WER)
<i>Baseline – Delta and Acceleration</i>	52.2	52.3
Cepstral bias adaptation	50.5	52.1

4 Trajectory Filtering

Trajectory filtering (see Figure 2) refers to a data transformation that is aimed at incorporating context information at long time scales. Several different sub-projects were pursued and details of the most promising approaches are found in Appendices 3 and 4. The results from this investigations were encouraging but experiments with the large test set on SWITCHBOARD did not yield word error reduction.

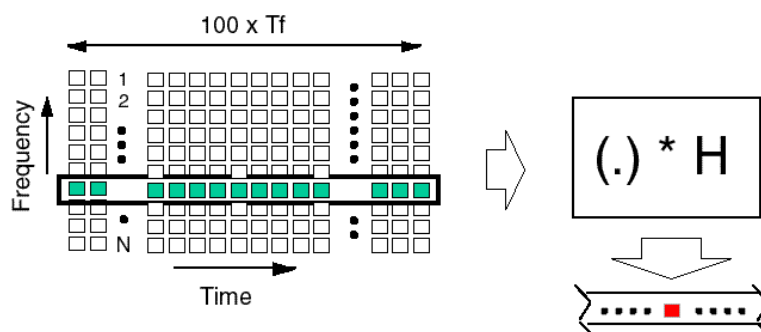


Figure 2: Trajectory Filtering

5 Optimal Feature Extraction

Optimal feature extraction using LDA (Linear Discriminant Analysis) and HDA (Heteroscedastic Discriminant Analysis) are depicted in Figure 3. The theoretical analysis and experimental details for HDA can be found in Appendix 5. This line of research has given very encouraging results not only on small scale experiments but also on the large training and testing set of WS-97. Results are summarized in the Ta-

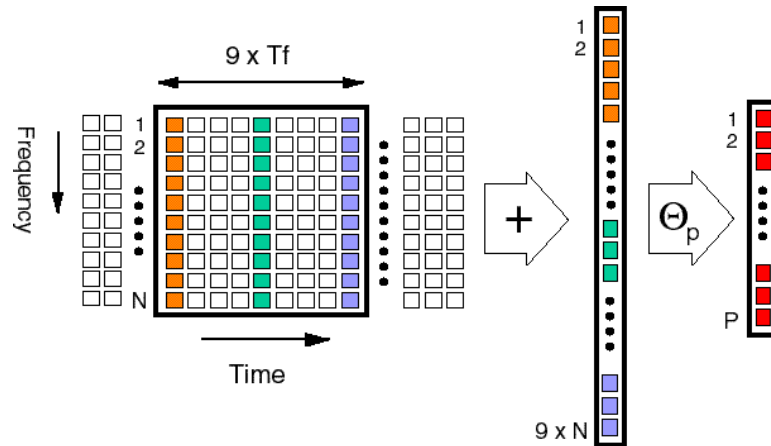


Figure 3: Linear and Heteroscedastic Discriminant Analysis

ble 2. Given that only just ONE experiment was performed using in this sub-project, the results warrant further investigation and government funding.

Table 2: Best recognition performance on LDA and HDA transformed Mel-cepstrum features and diagonal covariance matrix, unequal variance models.

Method Used	Feature Dimension	Recognition Error (%)
<i>Baseline – Delta and Acceleration</i>	39	49.9
LDA ($C = 4$)	39	51.1
HDA ($C = 4$)	39	49.1