DERIVATION OF TRAPS IN AUDITORY DOMAIN

Petr MOTLICEK, Doctoral Degree Programme (4)
Dept. of Computer Graphics and Multimedia, FIT, BUT
E-mail: motlicek@fit.vutbr.cz

Supervised by: Dr. Jan Černocký, Prof. Hynek Heřmanský

ABSTRACT

This contribution presents potential straightforward technique to extract temporal information in auditory domain. Even though the final phoneme accuracy is comparable to the traditional approach, it can be well suited to replace standard spectrum based techniques used in ASR systems due to higher flexibility and computational inexpensiveness.

1 INTRODUCTION

Most of feature extraction methods used in current Automatic Speech Recognition (ASR) systems are based on spectrum. However, such based techniques have distinct disadvantages, because they can be easily influenced by variety of issues, such as communication channel distortions or narrowband noise. Moreover, some other supplementary techniques need to be applied to deal with realistic communication environments.

Many of the noise-robust techniques employ the temporal domain processing operations to increase robustness in ASR. Psychoacoustic experiments prove that peripheral auditory system in humans integrates information of much larger time spans than the temporal duration of the frame used in traditional speech analysis. This time span is of the order of several hundred milliseconds. As the example of successive temporal domain based techniques are dynamic cepstral coefficients. These coefficients are computed as the first and second order orthogonal polynomial expansions of feature time trajectories, and are referred to as delta and acceleration coefficients, respectively. They represent the slope and curvature, respectively, of the feature trajectories, and are typically computed over 50 ms to 90 ms speech segments. Cepstral mean normalization, in which the long-term average is subtracted from the logarithmic speech spectrum, is another temporal processing technique.

Recently, many progressive temporal domain processing algorithms have appeared, where conventional spectral feature in phonetic classification is substituted by a several hundred millisecond long temporal vector of critical band energies [1]. The phonetic class is defined with respect to the center of this temporal vector. The stream of these vectors

goes to the input of classifier that attempts to capture the appropriate temporal pattern (TRAP), and is called TRAP classifier (Fig. 1).

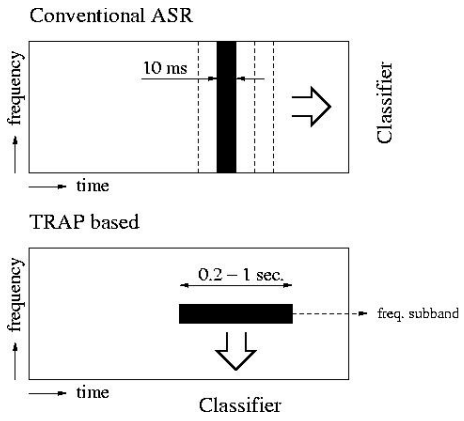

Figure 1: Innovative idea of employment temporal information in ASR.

In the original approach, a set of vectors representing its temporal evolution is extracted from a particular time trajectory. Critical bands are usually used as a basis of these time trajectories. In our approach we want to show that TRAPs do not have to be represented by time trajectories of spectral energy and can be derived a different way without applying any spectral processing operations.

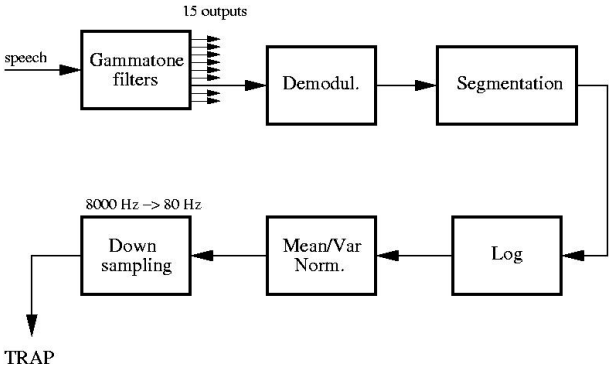

Figure 2: Derivation of TRAPs in auditory domain.

2 DERIVATION OF TEMPORAL PATTERNS

TRAPs are often examined in time evolution of basic sound units, phonemes, typically used in ASR. Traditionally, the speech signal is processed as a series of independent short-time (e.g. 10 ms) frames. Each frame is transformed into spectral domain using Fourier transform, and logarithmic critical band energies are derived.

In our approach, TRAPs are fully derived in auditory domain (Fig. 2). To preserve frequency independence of classification, some sort of band pass filter bank needs to be

applied. Such analysis filter bank is being represented by gammatone filters whose center frequencies and bandwidths match those of the critical bands. These linear phase gammatone filters are applied to the input signal to obtain an auditory-based time-frequency parametrization, which approximates the patterns of neural firing generated by the auditory nerve, and preserves the temporal information carried in speech.

Gammatone filters can be implemented using FIR or IIR filters [3]. In our approach, FIR filters were used in order to implement linear phase filters with the same delay in each critical band. The analysis filters have a length of $2N-1$ coefficients. They were obtained by convolving a sampled gammatone impulse response $g(n)$ of length $N = 128$ with its time reverse, where:

$$g(n) = a(nT)^{N-1} e^{-2\pi bERB(f_c)nT} cos(2\pi f_c nT + \varphi). \quad (1)$$

T is the sampling period, f_c is the center frequency, n is the discrete sample index, a, b are constants, and $ERB(f_c)$ is the equivalent rectangular bandwidth of an auditory filter. For an 8 kHz sampled speech, 15 FIR filters were used.

To extract the energy from each pass band filtered speech signal, the signal needs to be demodulated. Therefore filtered signal is multiplied by complex exponential $e^{j2\pi f_c nT}$, where j is the complex operator. Finally, low pass filter (LPF) is applied to preserve only non-modulated spectral components.

Our approach is not consistent with traditional method [1] in sense of derivation temporal patterns from logarithmic critical band energies. In spectral analysis based technique, the speech signal is processed as a stream of frames in order to capture non-stationary characteristic of the speech signal (the speech is downsampled according to frame length and frame shift, and frames are then used to derive final temporal trajectories). Due to derivation of TRAPs in auditory domain, the signal is still fully sampled, so that the length of extracted TRAPs (hundred milliseconds) is largely higher than length of originally derived TRAPs. The extraction of TRAPs from demodulated signals (each critical band is processed independently) is done the same way as traditional framing. The signal is divided into segments with some overlapping constant and the appropriate segment length. Each such segment, is Hamming windowed, processed by logarithm and the mean is subtracted. Created TRAPs have finally the same segment rate as in the original approach. Temporal evolution achieved by individual TRAP is sampled with primary sampling frequency, which is $F_s = 8\ kHz$ in our experiments. The spectrum that can be computed from temporal trajectory of critical band is referred to as modulation spectrum. Components of this spectrum for clean speech varies approximately in period of 1 to 20 Hz. Spectral components the vary more rapidly or slowly are caused by non-speech artifacts and do not carry any efficient information. Therefore we can downsample these temporal trajectories at least by ratio 200 (modified F_s will be 40 Hz) with appropriate low pass filtering. The whole previously described technique for derivation of TRAPs in auditory domain is given in Fig. 3.

3 EXPERIMENTAL SETUP

To get understanding of the information that is available in the time trajectories, we examine for patterns in the temporal evolution of phonemes. Therefore phoneme labeled

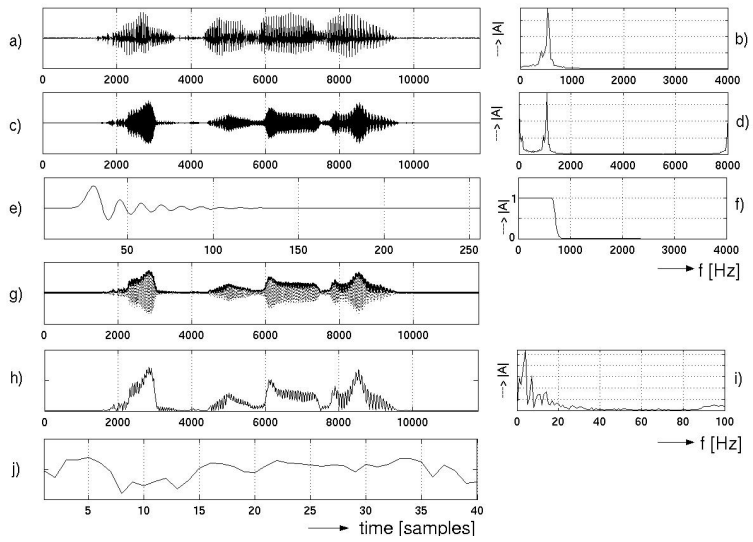

Figure 3: a) Input speech sentence ($F_s = 8$ kHz). b) Spectrum of input signal passed through 5^{th} gammatone band pass filter with $f_c = 531$ Hz. c) Time domain interpretation of filtered speech. d) Spectrum of filtered signal and multiplied by complex exponential. e) Impulse response of the following LPF. f) Amplitude frequency response of the following LPF. g) Filtered speech (dotted line) with demodulated energy (solid line). h) Energy extracted from band passed signal. i) Modulation spectrum related to the energy of band passed signal. j) Final TRAP - 1 sec. of energy after application of logarithm and mean normalization with downsampling ratio $R = 200$.

database is needed for our experiments [2]. Each single critical band is classified into phonetic classes by a multi-layer perceptron (MLP) with 3 layers. The size of input layer is determined by the length of TRAP. The hidden layer with sigmoid non-linearities have 300 neurons. The size of output layer is given by the number of classes. TIMIT database with 42 phonetic classes is used to train individual band classifiers. The training data is split into a training and cross-validation (CV) sets. Outputs of band classifiers are class posteriors that are gaussianized (application of logarithm). Since there are 15 critical bands available within the speech bandwidth, we have at our disposal 15 different TRAP outputs.

Then we use another MLP for combining the outputs obtained from each of the 15 TRAPs. The merger consists of 3 layers. The input to the combining network (called merger) is the concatenated vector of posteriors of the 42 phonetic classes from each of the 15 TRAPs (42×15). The hidden layer contains 300 neurons. The size of output layer is given by the number of classes (42). The merger is usually trained on different data than used for training band classifiers. We used OGI-stories corpus [1] and considered 42 phonetic classes (same as for TIMIT). Therefore, for this new training data, TRAPs must be generated and forward passed through band classifiers.

The phoneme recognition accuracy for a previously described classification in each critical band is in the range of 21% - 25%. Tab. 1 shows the final phoneme recognition accuracy of the merger on the CV and train set of OGI-stories corpus. It is related to the 500

ms long TRAPs, 12.5 ms frame shift that results into 40 samples of TRAP (downsampling ratio $R = 100$).

Technique	Best CV acc. [%]	Best train acc. [%]
Traditional	51.49	61.01
Our	50.68	62.53

Table 1: Performance with the TRAPs.

4 CONCLUSIONS

It has already been shown and published (and also successfully employed in feature extraction algorithm for ASR [4]) that information extracted from temporal trajectories can largely increase ASR performance, mainly when combined with classical features. However, the solely proposed technique was based on spectrum analysis for derivation of TRAPs. In this paper we gave a brief description of different technique for extraction temporal information employed in auditory domain. The final performance on CV subset is comparable (little bit worse for CV subset and better for train subset) to the traditional approach. These results also show that with reasonable larger train corpus we should be able to achieve higher final performance.

Proposed approach is advantageous in terms of possible modifications and computational inexpensiveness. For instance, it is effortless to change the time length of created temporal segments, without touching frame shift (just downsampling ratio is modified), and so on.

ACKNOWLEDGMENTS

This research has been partially supported by industrial grant from Qualcomm, DARPA N66001-00-2-8901/0006, by Grant Agency of Czech Republic under project No. 102/02/0124 and by EC project Multi-modal meeting manager (M4), No. IST-2001-34485.

REFERENCES

[1] Hermansky H., Sharma S.: TRAPS - Classifiers of Temporal Patterns, Proceedings of ICSLP'98, Sydney, Australia, November 1998.

[2] Černocký J.: TRAPS in all senses, Report of post-doc research internship in ASP Group, OGI-OHSU, September 2001.

[3] Gold B., Morgan N.: Speech and Audio Signal Processing, John Wiley & sons, inc., New York, 1999.

[4] Jain P., Hermansky H., Kingsbury B.: Distributed Speech Recognition Using Noise-Robust MFCC and Traps-Estimated Manner Features, Proceedings of ICSLP'02, Denver, USA, September 2002.