Time-domain based Temporal Processing with Aplication of Orthogonal Transformations

Petr Motlíček, Jan Černocký

Faculty of Information Technology, Brno University of Technology
Božetěchova 2, Brno, 612 66, Czech Republic
{motlicek,cernocky}@fit.vutbr.cz

Abstract

In the paper, novel approach that efficiently extracts the temporal information of speech has been proposed. This algorithm is fully employed in time-domain, and the preprocessing blocks are well justified by psychoacoustic studies. The achieved results show the different properties of proposed algorithm compared to the traditional approach. The algorithm is advantageous in terms of possible modifications and computational inexpensiveness. Then, in our experiments, we have focused on different representation of time trajectories. Classical methods that are efficient in conventional feature extraction approaches showed not to be suitable to approximate temporal trajectories of speech. However, the application of some orthogonal transformations, such as discrete Fourier transform or discrete cosine transform, on top of previously derived temporal trajectories outperforms classification in original domain. In addition, these transformed features are very efficient to reduce the dimensionality of data.

1. Introduction

In traditional ASR system, the speech signal is processed as a series of independent short-time (10 ms to 100 ms due to delta coefficients) frames in order to capture non-stationary characteristic of the speech signal and to facilitate application of the well-developed processing techniques for stationary signals. Spectral features are usually presented in form of filter bank energies, Linear prediction coefficients (LPCs), cepstral coefficients, Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) coefficients, and are the basis of the most feature extraction methods in current ASR. They describe the spectral envelope of the speech signal in a given frame. However, their disadvantage is its strong sensitivity to changes in the communication environment caused by different channel characteristics or background noise [4].

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 (around 200 ms). One evidence in auditory perception is the phenomenon of forward masking [1].

It has already been shown and published (and also successfully employed in feature extraction algorithms for ASR [2]) that information extracted from temporal trajectories can largely increase ASR performance, mainly when combined with classical features.

Recently, relatively long temporal trajectories of the speech features (0.5s - 1s) have been examined to capture phonetic information that is presented in time. This study has resulted into a set of new features (TRAPs) incorporating medium-time temporal dependency in the features used in the multi-band system. These TempoRAl Patterns (TRAPs) are derived from spectral energies obtained by standard spectral processing operations [4]. In our research, we propose novel algorithm attempting to extract temporal information of speech directly in time (auditory) domain. In other words we show that solely temporal processing based operations can be used to derive temporal patterns that generate the set of features. This proposed algorithm can be easily modified. For instance, it is effortless to change the time length of created temporal segments, without touching frame shift, and so on.

Then, our experiments were focused on representation of these time trajectories that resulted into an application of some well-known orthogonal transformation applied on top of TRAPs. Such TRAPs projected into different domain are supposed to be more accurate for classification and provide higher recognition performance. The valuable property of potential transformation is the data reduction without any system degradation. We have also experimented with importance of phase information, that capture TRAPs, from a classification point of view.

2. Experimental setup

In practical experiments we have used feed-forward multi-layer perceptron (MLP) based classifiers. The availability of information in time trajectories is evaluated on temporal evolution of phonemes. Therefore, phoneme labeled database is needed for our experiments. TIMIT database, prepared at the National Institute of Standards and Technology (NIST), with $N = 42$ phonetic classes is employed to train individual band classifiers. Details are given in [5].

TRAP features are built on multi-band approach. For each of 15 frequency bands, an MLP based classifier is used. It classifies TRAPs into phonetic classes. Each band classifier is a MLP with 3 layers. The size of input layer is determined by the length of TRAP. The hidden layer has in most of experiments 300 neurons. The size of output layer is given by the number of classes. For experimental purposes, the training data is split into training and cross-validation (CV) sets. Description and more details are given in [5].

After band classification we use another MLP (Merger) for combining the outputs obtained from each of the 15 TRAPs. Merger is trained on OGI-Stories corpus. Full description is given in [6]. The merger consists of 3 layers. The input is the concatenated vector of posteriors of N phonetic classes from each of the 15 TRAPs ($N \times 15$). The hidden layer contains 300 neurons.The size of output layer is given by the number of

classes (N). The merger is trained on data different from those used for training band classifiers. Therefore, for this new training data, TRAPs must be generated and forward passed through band classifiers.

Phoneme recognition accuracies were used to analyze the performance of MLP based classifiers. The evaluation of performance was based on results seen during the training and cross-validation of nets. The following results are provided:

- final phoneme recognition accuracy on the CV sets of TIMIT database while training the band classifiers for 3 bands (0^{th}, 5^{th},10^{th}) on TIMIT (noted as TCV_0, TCV_5, TCV_{10}).
- phoneme recognition accuracy for OGI-Stories forward passed through the band classifiers for 3 bands (0^{th}, 5^{th},10^{th}), (noted as SFW_0, SFW_5, SFW_{10}).
- final phoneme recognition accuracy on the cross-validation set (noted as FCV) and training set (noted as FT) of OGI-Stories in merger training.

3. Original approach to TRAP-derivation

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. These energies have been initially used to provide temporal information of speech [4].

As mentioned above, the temporal structures of phonemes were analyzed to get understanding the nature of the linguistic information available in the temporal structure of speech. For given phoneme label, all segments from particular frequency band are extracted. Then, for each such segment, several hundred milliseconds long TRAPs are formed that are centered around each frame of the segment labeled as the phoneme.

Traditionally derived set (noted as TR_{traps}) of 41-point TRAPs (500 ms), derived from critical band logarithmic spectral energies[4], with TRAP-based mean normalization, and with the tandem of 15 band-classifiers and the merger, gives the performances mentioned in Tab. 1.

4. Derivation of temporal patterns in time-domain

In our novel approach we show that TRAPs do not have to represented by time trajectories of spectral energies and can be fully derived in time-domain without applying any spectral processing operations. The whole technique is shown in Fig. 1.

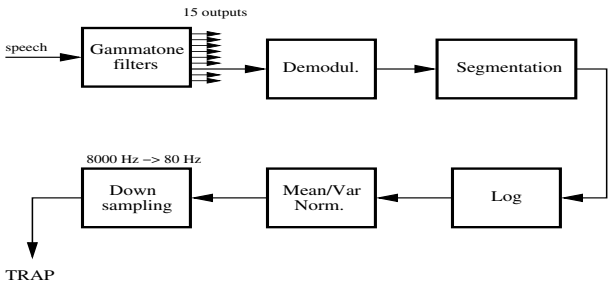

Figure 1: Derivation of TRAPs in time-domain.

To preserve frequency independence of classification, a band pass filter bank needs to be applied. In our approach, such analysis filter bank is represented by gammatone filters [7], whose center frequencies f_{c_i} 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.

The speech signal, filtered by the bank of band pass filters, results into 15 trajectories with the spectrum shifted in frequency by f_{c_i}. The extraction of the energy from each band pass filtered speech signal is equivalent to the shifting that signal's spectrum down in frequency by f_{c_i}. It can be done by multiplication a time signal by a single complex exponential $e^{-j2\pi f_{c_i} nT}$, where j is the complex operator. This operation is usually called "complex downconversion". Finally, low pass filter (LPF) is applied to preserve only non-modulated spectral components. A cutoff frequency of LPF is dependent on previously applied band pass gammatone filters, because of their frequency-varying bandwidth. The extraction of TRAPs from demodulated signals 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. Such derived temporal trajectories are still fully sampled, so that the length (in samples) of extracted TRAPs (hundred milliseconds) is largely higher than length of originally derived TRAPs. However, due to the properties of modulation spectrum of the speech [8], these temporal trajectories can be downsampled without loosing any information important for their classification. The final sampling frequency is $F_{sampl} = 80\ Hz$.

The results with 40-point TRAPs (related to 500 ms time length, downsampling ratio $R = 100$) derived in time-domain are in Tab. 1 (noted as AD_{traps}).

5. Representation of temporal patterns

Our initial experiments attempted to apply well-known algorithms successfully used in standard feature extraction, that are able to reduce useful information of speech. One of them is Linear prediction, where the power spectrum of speech is approximated by autoregressive model [7]. The use of such algorithm leads into drop-out of the phase information that is irrelevant in static feature extraction. However, initial experiments with TRAPs, approximated by autoregressive model, showed that phase information is very important for their classification. Therefore, operations that preserve also phase information of input data need to be employed. In our experiments we have attempted to use Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), Discrete Hadamard Transform (DHT) and discrete Karhunen-Loève transform (KLT). All these transforms were applied on TRAPs derived in time-domain.

5.1. Discrete Fourier Transform (DFT)

For the wide-sense stationary signals, DFT asymptotically approaches the eigen-decomposition. The existence of Fast Fourier Transform (FFT) algorithm for the computation of the DFT, and data independent nature of this transform are among its attractive features. For derivation of DFT, a periodic sequence $x[n]$ with period N is considered, so that $x[n] = x[n + rN]$. Such sequence can be represented by a Fourier series (as with continuous time periodic signals), correspond-

Technique	TCV_0	TCV_5	TCV_{10}	SFW_0	SFW_5	SFW_{10}	FT	**FCV**
TR_{traps}	25.74	22.9	22.61	25.48	25.48	22.07	61.01	**51.49**
AD_{traps}	20.93	24.14	23.50	20.30	25.00	24.04	62.53	**50.68**

Table 1: Phoneme recognition accuracies (in %) of the MLP based classifiers.

ing to a sum of harmonically related complex exponential sequences. Initially, the magnitude modulation spectrum has been used for MLP classification. Each 40 samples long TRAP was zero padded and transformed using DFT. First 40 components of magnitude modulation spectrum were used for MLP classification. Zero padding was applied in order to obtain 40 sampled halves of magnitude modulation spectrum. The results are given in Tab. 2 (FFT_{abs} experiment), and show poorly trained MLP classifiers. In case of FFT_{a+p} experiment, where an additional sequence of phase coefficients was used with magnitude spectrum components, large improvement can be observed. Here, the size of input vectors for MLP based band classifiers is also 40 (20 magnitude spectrum components, 20 phase spectrum components). Even though, features used for classification contain the same information as baseline features (AD_{traps} experiment), the final recognition performance is still much worse than the baseline. Much better results are achieved, when real and imaginary components of modulation spectrum are employed in MLP classification (FFT_{r+i} experiment). Such trained MLP classifier outperforms the baseline system.

5.2. Discrete Cosine Transform (DCT)

In case of the DFT, the basis sequences are the complex periodic sequences $e^{j2\pi kn/N}$, and the resulting sequence $X[k]$ is, in general, complex even if the input sequence $x[n]$ is real. It is natural to inquire as to whether there exist sets of real-valued basis sequences that would yield a real-valued transform sequence $X[k]$ when $x[n]$ is real. This has led to the definition of a number of other orthogonal transform representations. One of them is the discrete cosine transform (DCT). In case of all definitions of DCT, even periodic sequences need to be created from finite-length sequences. Generally, there are four common ways to create such sequences which result into four forms of DCT [9]. In first experiments, DCT form that is easily derived from DFT has been used. One period of resulting sequence $\tilde{x}[n]$ can be written as:

$$\tilde{x}[n] = x[((1))_{2N+1}] + x[((n+1))_{2N+1}] + x[((-n-1))_{2N+1}],$$

where N is the length of original sequence $x[n]$, and $((.))_{2N+1}$ stands for zero padding the sequence to the length of $2N+1$. For such resulting periodic sequence $\tilde{x}[n]$, standard DFT can be used (instead of any traditionally defined DCT form). DCT sequence for each TRAP was computed. In order to get the understanding of the importance of particular DCT coefficients, two following experiments were run:

- First, full DCT spectrum was used for classification ($0 - F_{sampl}/2$), where $F_{sampl} = 80\ Hz$. Then, DCT spectrum was reduced using different cut off frequency F_{cutoff}, and just low part of reduced spectrum was used for classification. The original TRAP sequence $x[n]$ was possibly zero padded, before its DCT projection, to obtain constantly 40 final DCTs for classification. Therefore, the size of input layer of MLP classifier was constant for all the experiments, so that it did not affect the performances. The experimental results for different F_{cutoff} are given in Tab. 2, and graphically shown in Fig. 2. It shows that for $F_{cutoff} \gtrsim 16\ Hz$, the recognition performance almost does not change. However, when lower part of DCT spectrum than approximately $16\ Hz$ is cut off, the system performance dramatically degrades.
- On contrary to the first set of experiments, next, the DCT based spectrum was limited from lower to upper frequencies. The highest frequency was kept constant ($40\ Hz$), whereas the lowest cut off frequency F_{cutoff} changed in interval $0 - 20\ Hz$. The results are shown in Fig. 3.

As mentioned above, the DCT corresponds to forming a periodic, symmetric sequence from a finite-length sequence in such a way that the original finite-length sequence can be uniquely recovered. There are many ways to do this, therefore, there are many definitions of DCT. In first case, the DCT based on DFT has been used. In other experiments DCT1 and DCT2 according to [9] were used.

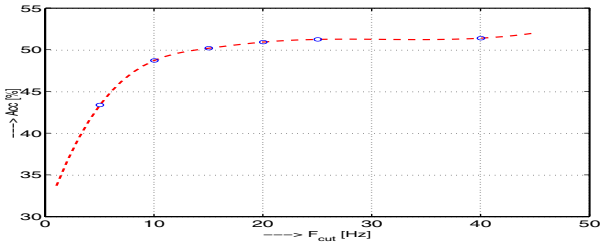

Figure 2: Dependence of phoneme recognition accuracy on F_{cutoff}, when the upper spectral components are discarded.

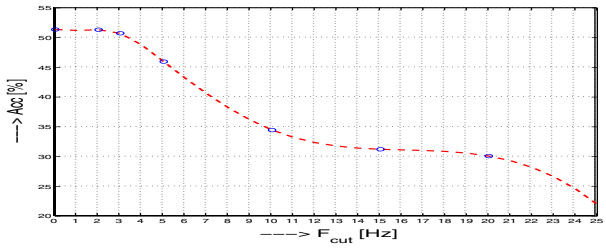

Figure 3: Dependence of phoneme recognition accuracy on F_{cutoff}, when the lower spectral components are discarded.

5.3. Discrete Karhunen-Loève transform (KLT)

The another linear orthogonal transform that is possible to use is the KLT. It is widely employed in signal processing, statistics, and neural computing. In some application areas, it is also called the Principal Component Analysis (PCA) [3]. Principal components (basis of KLT) are derived from covariance matrix

Technique	TCV_0	TCV_5	TCV_{10}	SFW_0	SFW_5	SFW_{10}	FT	**FCV**
FFT_{abs}	16.21	17.26	17.91	16.51	21.95	21.91	46.26	**35.91**
FFT_{a+p}	17.67	20.54	20.41	17.56	23.25	22.79	55.92	**45.44**
FFT_{r+i}	21.41	25.03	23.83	18.93	25.33	24.02	62.39	**51.73**
DCT	22.23	24.93	24.04	19.86	25.23	23.87	61.97	**51.37**
$DCT1$	22.31	25.27	23.86	19.32	25.30	23.78	62.04	**51.61**
$DCT2$	22.61	24.75	23.93	20.1	25.24	24.03	62.38	**51.48**
KLT	22.46	25.13	23.95	20.23	25.49	23.76	62.79	**51.54**
DHT	21.71	25.27	24.39	18.74	25.24	24.24	62.68	**51.76**

Table 2: Phoneme recognition accuracies (in %) of the MLP based classifiers.

that is obtained from the data. TIMIT training data were employed to obtain principal components, and then KLT was applied on TIMIT as well as on OGI-Stories. No dimensionality reduction was performed in order to keep the size of input layer of MLP band classifiers unchanged. Therefore, each 40 samples long TRAP was projected into 40 principal components (compounded according to the highest eigen values of eigen vectors). The achieved phoneme accuracies are given in Tab. 2.

5.4. Discrete Hadamard transformation (DHT)

Unlike the other well-known transforms, such as the DFT and DCT, the elements of the basis vectors of the DHT take only the binary values $+1$ and -1. Therefore, they are well suited for digital signal processing applications where computational simplicity is required. The basis vectors of the *m*-point DHT can be generated by sampling a class of functions called Walsh functions. The one dimensional DHT of a sequence $x[n]$ is defined as:

$$X[k] = \sum_{n=0}^{N-1} x[n](-1)^{\sum_{i=0}^{m-1} b_i(m)b_i(k)}, \quad 0 \leq k \leq N-1,$$

with $N = 2^m$, where *m* is an integer. $b_i(z)$ is the i^{th} bit in the binary representation of z. The results of application of DHT are given in Tab. 2.

6. Conclusions

From achieved results, mentioned in Tab. 1, we conclude that band classifiers trained on traditionally derived TRAPs (TR_{traps} experiment) give better performances especially for low frequency bands than those trained on time-domain based TRAPs (AD_{traps} experiment). For higher bands the differences between TR_{traps} and AD_{traps} are minimal. The final phoneme recognition accuracies FT and FCV give us the ultimate number that can be used for comparisons of other systems. FT shows the recognition performance of system on data used for training, whereas FCV is associated with cross-validation data. TR_{traps} system yields slightly higher performance for FCV than our proposed system AD_{traps}. However, in case of FT, we obtain substantially better phoneme recognition accuracy for AD_{traps}. These results show the different behavior of these two algorithms. TR_{traps} seems to be more robust to the unseen data, whereas with AD_{traps} based technique we are supposed to obtain higher recognition performance with a well trained classifier (availability of large amount of training data). Moreover, proposed approach is advantageous in terms of possible modifications and computational inexpensiveness.

The following experiments (Tab. 2) definitely show that the phase information of temporal trajectories is important for their classification. Classical algorithms that are efficient in static feature extraction are not suitable in temporal domain processing. The results also show that DFT, DCT, KLT, and DHT applied on top of TRAPs slightly outperform classification in original domain. In addition, these features are very efficient in data reduction so that less complex classifiers can be used. Eventually, they generalize better than other conventional features and yield considerable complementary information with respect to short-term cepstral features in ASR.

7. Acknowledgments

This research has been partially supported 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. Jan Cernocky has been also supported by a post-doctoral grant from the Grant Agency of Czech Republic, no. 102/02/D108.

8. References

[1] B. C. J. Moore. “An Introduction to the Psychology of Hearing”, 3rd ed., Academic Press, New York/London, 1989.

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

[3] I.T. Jolliffe. “Principal Component Analysis”. Springer-Verlag, 1986.

[4] H. Hermansky, S. Sharma. “TRAPS - Classifiers of Temporal Patterns”, Proceedings of ICSLP’98, Sydney, Australia, November 1998.

[5] J. Černocký. “TRAPS in all senses”, Report of post-doc research internship in ASP Group, OGI-OHSU, http://www.fit.vutbr.cz/∼cernocky/publi/2001/report.pdf, September 2001.

[6] S. Sharma, “Multi-Stream Approach To Robust Speech Recognition”, PhD thesis, OGI, Portland, USA, April 1999.

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

[8] N. Kanedera and H. Hermansky and T. Arai. “Desired characteristics of modulation spectrum for robust automatic speech recognition”, Proceedings of ICASSP’98, vol. 2, pp. 613-616, Seattle WA, USA, 1998,

[9] A. V. Oppenheim, R. W. Schafer. “Discrete-Time Signal Processing”, 2nd Ed., Prentice-Hall, NJ, USA, 1998.