Localized Spectro-Temporal Features for Noise-Robust Speech Recognition

Gy. Kovács* and L. Tóth*
* Reseach Group on Artificial Intelligence, University of Szeged and Hungarian Academy of Sciences
Szeged, Hungary
Kovacs.Gyorgy.4@stud.u-szeged.hu, tothl@inf.u-szeged.hu

Abstract—**In speech recognition there has been a trend to incorporate more and more knowledge about human hearing into the feature extraction step. One such approach is the application of localized spectro-temporal analysis, which is inspired by neurophysiological studies. Here we experiment with extracting features from the patches of the widely used criticial-band log-energy spectrum by applying the two-dimensional cosine transform. Compared to earlier similar studies with the spectrogram representation, we find that our method is not worse, and faster. In experiments with noisy speech the proposed representation proves more noise-robust than the conventional mel-frequency cepstral features.**

I. Introduction

The traditional feature extraction methods used in speech recognition have a mathematical basis, and take only the most fundamental properties of articulation and hearing into consideration. For example, the most commonly applied mel-frequency cepstral coefficients (MFCC) [1] smooth out the fine details of the spectrum, as it is known that phonetic information is carried mainly by the spectral envelope. Also, it warps the linear frequency scale to the quasi-logarithmic mel scale, which is known to fit human hearing better. But in other respects it is just a mathematical tool based on conventional signal processing algorithms such as the Fourier transform and the cosine transform. Although it is not strictly necessary that processing methods which seek to mimic human hearing should outperform the purely mathematical algorithms, in general it seems reasonable to expect a better behavior from the methods that approximate the properties of human hearing more closely. One such property is the joint spectro-temporal sensitivity of the receptive fields of cortical cells [2]. Compared to what is known about the time-frequency tuning of these cells, the resolution of the conventional MFCC representation is much narrower in time and much wider in frequency. In this paper we experiment with a quite simple modification of the MFCC algorithm that works with localized spectro-temporal patches of the spectral representation instead of the narrow time-span and global frequency-span windows of the standard MFCC technique. The main purpose of these experiments is to show that the proposed localized spectro-temporal features are no worse than the conventional feature set when applied in phonetic classification.

In addition to the neurophysiological and psycho-acoustic findings, there is a further, purely practical argument for applying a localized feature extraction method: when the signal is corrupted with band-limited noise, a spectrally global analysis technique such as the MFCC will result in *all* the features being contaminated by the noise. When using localized patches, however, only a subset of the features will be affected, and therefore this approach should be more robust to noise. To test this hypothesis, a second set of experiments will be presented that compares the performance of the conventional and the proposed representation with varying noise levels.

II. Conventional Feature Extraction

All feature extraction methods seek to perform some sort of spectral analysis of the speech signal. As the signal is continuously changing, it is normally sliced up into small, uniform-sized pieces ('frames') of 20-30 milliseconds, during which the signal can be considered quasi-stationary. Then some spectral analysis is performed over these frames, typically using widespread signal processing tools such as the fast Fourier-transform (FFT). Though more sophisticated tools, for example filter banks adjusted to the sensitivity of the human ear may also be applied [3], here we will try to keep the computational complexity low, and we will use the simple FFT-based spectrogram as a starting point. This is also the first step of MFCC computation [1].

In some of the experiments we will extract the time-frequency patches directly from the spectrogram. We, however, would like to make our feature extraction as similar to the MFCC calculation as possible. The second step of MFCC extraction is the warping of the frequency scale. This is motivated by the fact that the human ear has a sensitivity that gradually decreases at higher frequencies, in contrast to the linear resolution provided by the FFT. The MFCC algorithm simulates this by summing bands of the Fourier spectrum using triangular-shaped weighting functions that are placed uniformly along the auditory-motivated mel-scale and which have a shape that is a coarse approximation of the critical bands of human hearing. The typical resolution of this representation is 100 feature vectors (i.e. columns) per second along the time axis and 25-50 spectral bands (i.e. rows) along the frequency axis. The energy values of both the conventional spectrogram and the critical-band energy spectrum are usually displayed on a decibel (i.e. logarithmic) scale. Figs. 1 and 2 show a comparison of the resulting mel-scaled critical-band energy spectrum with the original, linearly scaled spectrogram. Clearly, the former has a

This research was partially supported by the TÁMOP-4.2.2/08/1/2008-0008 program of the Hungarian National Development Agency.

978-1-4244-7431-8/10/$26.00 ©2010 IEEE

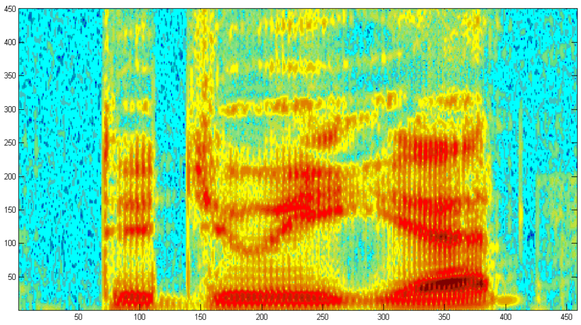

Figure 1. The classic spectrogram representation of a sentence.

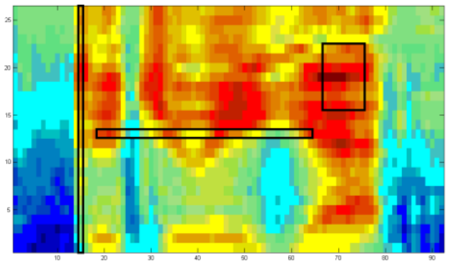

Figure 2. The critical-band energy representation of the sentence of Fig 1. The black boxes (left to right) show the shape of the feature extraction patches used by a) the classic MFCCs b) the TRAP features c) localized spectro-temporal features.

much coarser resolution, and it emphasizes the lower frequencies while the high-frequency parts are suppressed.

The final step of MFCC extraction is the application of the discrete cosine transform (DCT) to each spectral slice; that is, each column of the critical-band energy plot. This smooths the unimportant details of the spectral curve and retains only the envelope profile, and at the same time it also has an additional decorrelating effect. As smoothing is actually achieved by keeping just the first 10-15 DCT coefficients, this step inherently performs a dimensionality reduction as well.

One column of either the classic or the critical-band spectrogram gives a static picture of a very small time instant of the speech signal. However, many psychoacoustic experiments demonstrate that for the recognition of phones in fluent speech the dynamics of the spectrum is more important [4]. In speech technology this observation led to the introduction of the so-called delta features, which are derivative-like values extracted using a couple of earlier and later values of each static feature. For example, the HTK software package [5] we are going to use in the experiments applies the formula

$$\Delta_t = \frac{\sum_{j=1}^{\Theta} j(c_{t+j} - c_{t-j})}{2\sum_{j=1}^{\Theta} j^2}, \qquad (1)$$

where the resulting Δ_t's are the delta coefficients corresponding to the c static coefficient at time index t, and the default value of the window size Θ is 2.

Usually the same computation is repeated on each Δ, leading to the set of $\Delta\Delta$ values, and the concatenation of the static MFCCs, the Δ's and the $\Delta\Delta$'s yields the classic feature set of speech recognition.

III. Localized Spectro-Temporal Features

The method of processing the speech signal in uniform 20-30 millisecond chunks has its roots in the speech coding tradition, and is retained mostly for technical convenience. Humans can barely recognize such short speech excerpts, which clearly shows that they are not an optimal choice for the basic unit of classification. Though the Δ and $\Delta\Delta$ features capture pieces of information from the neighboring 4-4 frames, both physiological and psychoacoustic experimental results indicate that the human brain extracts information from even longer time spans, processing frequency bands quasi-separately [2, 6]. Based on these findings, several modifications have been proposed for the feature extraction step, which can be viewed as changing the size and shape of the time-frequency patch extracted from the spectral map (we illustrate some of these in Fig. 1). The simplest idea is to work with larger windows along the time-axis: in neural-net based recognizers it is now standard practice to train the system on 9 neighboring MFCC vectors [7]. The main drawback of this method is that the number of features grows significantly, and hence the application of dimension reduction methods such as principal component analysis (PCA) and linear discriminant analysis (LDA) may be necessary. Also, these features are still global along the frequency axis, while several psychoacoustic studies suggest that the windows should be localized in frequency as well. These studies motivated the introduction of the TRAP model by Hermansky et al. In this scheme each frequency band is processed separately, and the corresponding results are combined only at a later step; the time-span of the processed trajectory patterns even goes up to 1 second in certain experiments [8]. This setup can be interpreted as a sort of 'inverse' arrangement compared to the classic MFCC windows (see Fig. 1 again).

An obvious generalization is when the patches are localized in both time and frequency [9]. Here we are going to work with patches like this by following the study of Bouvrie et al. [10], and apply two-dimensional DCT to process the localized time-frequency patches extracted from the spectral map (cf. Fig. 1). Though the studies of human perception can be indicative of the proper size of these time-frequency patches, in machine learning experiments different values may result in optimal recognition performance, due to the various properties and peculiarities of the signal processing and machine learning algorithms applied. Thus the best we can do is to vary the sizes and look for the optimal parameters by empirical evaluation. A detailed description of these experiments are presented in the following two sections.

IV. Experiments with Clean Speech

All the experiments we report are phone classification experiments on the well-known TIMIT speech corpus [11]. In the train-test partitioning of the data we followed the widely accepted standard: the full set of the 3696 train sentences were used for training, and testing was always executed on the full test dataset of 1344 utterances. The train and test sets consisted of 142910 and 51681 phone instances, respectively. The phonetic labels of the database were fused into 39 categories, which is again standard practice [12].

In all the experiments a multi-layer perceptron neural net [13] was applied as a classifier. It contained one hidden layer of 500 neurons; the output layer applied the softmax nonlinearity, while the hidden neurons worked with the sigmoid function. The number of output neurons was set to the number of classes (39), while the number of inputs varied, as will be described later. The neural net was trained using standard backpropagation on 90% of the training data in semi-batch mode, and cross-validation on the remaining 10% was used as the stopping criterion. The cross-validation data was of course selected randomly.

A. Experiments with the conventional spectrogram

In the first set of experiments we sought to reproduce the results of Bouvrie et al. [10]. They use the conventional spectrogram as the time-frequency representation from which the localized feature patches are extracted. The spectrogram is obtained with the following parameters: the signal was cut into frames with 32 sample hops applying the Hamming window, and each frame was Fourier-transformed using a 1024-point FFT. We tried two configurations for the frame size, namely 300 samples were used for the narrow- and 150 for the wide-band cases. Then the log-magnitude of the resulting spectrum was taken, and it was normalized so as to have unit variance and a mean of zero for each utterance.

The next step is the extraction of the time-frequency patches, which was done by computing a sliding localized two-dimensional DCT over the spectrogram. The window and step sizes were again chosen based on the suggestions in [10] (with a slight modification – substituting even window sizes by odd ones – for symmetry reasons). Namely, 51 by 21 bin windows were applied in the narrow band case, and 41 by 51 bin windows in the wide-band case (always giving the height first). The step sizes were 25 bin for the frequency, and 2 bin for the time axis in both cases. In contrast to [10], the 51 by 21 bin window size was tested not only with the narrow-band, but with the wide-band spectral representation as well.

By default, the DCT returns the same number of coefficients as the size of its input array. Similar to the 1D case – that is, the computation of MFCC – one can throw away the coefficients which correspond to higher modulation frequencies. With this step we smooth out the unnecessary fine details from the spectrum and reduce feature dimensionality at the same time. Bouvrie et al. propose keeping only the 6 lowest-order 2D-DCT coefficients corresponding to the upper left 3x3 triangle of the coefficient matrix [10]. Apart from this configuration, in certain cases we also experimented with retaining more (9 or 15) coefficients in a similar manner.

Next, for the phone classification experiments we had to form a fixed-length feature vector from the variable number of 2D-DCT coefficients extracted from the patches belonging to each phonetic segment. For this purpose we followed the technique proposed by Halberstadt [14], which was also found to work well by other authors [15, 16]. Each phonetic segment was divided into three parts along the time axis, and the coefficients belonging to the same patch index and DCT coefficient index were averaged over time within these segments. Two further segments were composed from the 30-30 milliseconds of the signal before and after the segments, and were processed in a similar way. This technique yields a pooled feature vector that consists of the same number of components for each segment – five times the number of patches along the frequency axis and the number of DCT coefficients retained – independent of the segment duration. After, the segment duration was also appended to this segmental feature vector.

TABLE I.
PHONE CLASSIFICATION ERROR RATES USING THE SPECTROGRAM

Spectrogram resolution	2D-DCT patch size	No. DCT coeffs	No. features	Error rate
Narrow-band	51x21	6	511	20.53%
Wide-band	51x21	6	511	20.14%
Wide-band	41x51	6	511	20.66%
MFCC			196	20.30%

The results obtained with various patch sizes are shown in Table I. As a baseline result, the score obtained with the conventional MFCC features is also presented; these were extracted using the HCopy module of the HTK toolkit [5]. With the default parameters, 13 mel-cepstral coefficients were calculated over 25 ms time frames every 10 ms, and the vectors were augmented with the Δ and ΔΔ coefficients, as described in Section 2. The 39-component feature vectors were converted into a segmental feature vector of fixed size using the same averaging method outlined above. Including duration, each phonetic segment was represented by 196 features.

As the results show, the proposed features are quite insensitive to the exact parameters (resolution and patch size). All three scores are similar to the MFCC result, and the best one even slightly outperforms it. Because of the big patch sizes and the larger number of features, however, the extraction of the 2D-DCT features takes much longer than that of the conventional MFCCs. We will address this problem in the next subsection.

We should mention here that all our results are consistently better than those presented in [10], both with the conventional and the localized spectro-temporal features. We attribute this to the fact that though the processing of the patches was similar, we applied a different type of classifier.

B. Experiments with the critical-band energy map

The resolution of the spectrogram used by Bouvrie et al. is much higher than that usually applied in speech recognition, both in time and frequency. Bringing it closer to the conventional resolution may yield a reduction in computational cost, and also make a comparison with the standard features more believable. Moreover, nowadays it is widely accepted that the mel-warping of the frequency scale is useful for recognition, and so feature extraction methods that work on the linear frequency scale have mostly been abandoned. Motivated by these facts we decided to repeat the experiments on the same critical-band energy spectral representation that is used as the starting point of the MFCC computation, as described in Section 2. Fortunately, the HCopy module of HTK can be parametrized so that it calculates just the critical-band spectrum and skips the final step, the DCT computation. This way we could ensure that we worked on exactly the same spectral representation as the one from which the baseline MFCCs are obtained.

In the first pilot studies we adjusted the window size and step of the spectral computation so that it agreed with the wide-band resolution used in the spectrogram-based experiments. That is, we did not decrease the resolution of the spectrum, but only activated the mel-scale frequency axis warping. More precisely, 104-105 frequency bands

TABLE II.
PHONE CLASSIFICATION RESULTS WITH THE CRITICAL-BAND ENERGY MAP

Spectral resolution	No. channels	2D-DCT patch size	Patch step (horizontal)	Patch step (vertical)	No. DCT coefficients	No. features	Error rate
WB	105	17x21	6	2	6	511	19.55%
WB	104	15x21	6	2	6	511	19.90%
MFCC	105	17x5	6	1	6	511	20.71%
MFCC	105	17x7	6	1	6	511	20.65%
MFCC	105	17x9	6	1	6	511	20.31%
MFCC	105	17x11	6	1	6	511	21.01%
MFCC	50	9x9	3	1	6	481	20.60%
MFCC	50	15x9	3	1	6	451	21.01%
MFCC	52	7x7	3	1	6	511	20.53%
MFCC	52	7x9	3	1	6	511	20.77%
MFCC	53	9x7	3	1	6	511	20.22%
MFCC	53	9x9	3	1	6	511	20.32%
MFCC	26	5x9	2	1	6	361	20.43%
MFCC	26	5x15	2	1	6	361	22.45%
MFCC	26	7x9	2	1	6	361	20.27%
MFCC	26	7x15	2	1	6	361	22.39%
MFCC	26	5x9	2	1	9	541	20.04%
MFCC	26	7x9	2	1	9	541	19.88%
MFCC	14	7x9	2	1	9	271	20.66%
MFCC	14	7x9	2	1	15	451	19.73%
MFCC	13	5x9	2	1	6	181	21.87%
MFCC	13	5x9	2	1	9	271	20.37%
MFCC	13	5x9	2	1	15	451	19.79%

were extracted from the spectrogram (the number varies slightly in order to support the full coverage of the spectral bands by the patches). This results in a smaller spectral map height than that of the spectrogram, so the patch size and the patch hop along the frequency axis were proportionally decreased to 15-17 and 6, respectively. The error rates obtained with two different patch sizes are shown in the first two rows of Table II. Inspiringly, both scores are better than the earlier results.

The next step was to reduce the spectral resolution so that it was equivalent with that used for the MFCC computation. Applying the default MFCC settings, 100 critical-band energy vectors were extracted per second from 25 ms frames. As the number of critical-band frequency channels was set to 26 during the MFCC extraction, we mainly experimented with this many channels, but other experiments were also carried out where the number of channels was four times or two times higher, or just the half (104, 52, and 13 – again 1-2 channels were sometimes added to make the patches fit the full range). As both the time and frequency resolution of the spectral representation became much smaller than that of the spectrogram used in the previous subsection, the size of the time-frequency patches was also shrunk proportionally: for example, to 5 by 9 bins for the default 26 frequency channel case. The patch step size was 2 along the frequency axis and 1 along the time axis in this case, but, naturally, the proper step size again depends on the number of channels.

Lots of combinations of the number of channels, patch size and patch step were tried, and the results obtained with the various parameter settings are presented in Table II. As can be seen, the best scores were again slightly better than the results obtained with the spectrogram. Apart from some 15-long patches that performed significantly worse, the results are quite similar, independently of the actual settings. Hence the most important finding of these experiments was that the resolution of the input spectrum and the size of the DCT patches can be much smaller, therefore the computational costs can be decreased without losing recognition accuracy. Moreover, conventional tools such as the HCopy routine of HTK can be used for the critical-band log-energy extraction step.

V. NOISY SPEECH EXPERIMENTS

As the time-frequency window used by the conventional MFCC extraction contains all the frequency bins of a given time instance, the corruption of just a few frequency bands by noise will ruin the values of all the coefficients extracted. With localized time-frequency patches, however, only a subset of the patches gets corrupted by the noise, and hence only a subset of the features will be affected. Consequently, one can reasonably expect that the proposed features are less sensitive to band-limited noise than the conventional MFCCs. To test this, we artificially contaminated the test dataset with pink noise of various levels. The spectral distribution of pink noise has the highest energy at 0 Hz, and gradually tails off at higher frequencies, so the patches at higher frequencies are less and less contaminated. The amplitude of the noise was tuned to get a signal-to-noise ratio of 20, 10 and 0 decibels in three different experimental settings. The noisy dataset was provided by the authors of [10]. We should again emphasize that in all the noisy experiments *training was performed on the clean data* and only the testing was executed on the noisy dataset.

TABLE III.
PHONE CLASSIFICATION ERROR RATES WITH NOISE-CONTAMINATED TEST DATA

Feature set	Clean	Pink noise			Babble noise		
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB
MFCC	20.30%	36.76%	63.00%	80.13%	31.58%	55.23%	77.46%
2D-DCT on wide-band spectrogram	20.66%	31.15%	49.74%	72.70%	31.03%	52.59%	76.04%
2D-DCT on critical bands 53 chans, 9x7 patches, 6 coeffs	20.22%	37.83%	60.35%	80.30%	30.93%	51.69%	76.84%
2D-DCT on critical bands 26 chans, 7x9 patches, 9 coeffs	19.88%	35.65%	58.86%	78.69%	33.05%	52.59%	75.82%
2D-DCT on critical bands 13 chans, 5x9 patches, 15 coeffs	19.79%	34.83%	57.75%	77.17%	35.06%	55.43%	77.25%

For the noisy tests we chose three configurations from the clean data experiments with 53, 26 and 13 critical bands. Of course, the MFCC tests were also repeated on the noisy test data to have a comparative baseline. The results obtained for pink noise are shown in Table III. As can be seen, the 2D-DCT features yielded lower error rates than the MFCCs in almost every case, especially with higher levels of noise. It also turned out, however, that the feature set extracted from the spectrogram behaved much better than the critical bands-based set. The reason might simply be the unlucky choice of the noise type: pink noise affects the higher frequencies less, while these frequencies are over-represented in the linear frequency-scale spectrogram compared to the critical-band energy map due to the mel-warped frequency scale of the latter. To justify this hypothesis we repeated the experiments with babble noise, which simulates the effect of several people talking in the background. The babble noise sample was taken from the NOISEX-92 database [17], and the FaNT tool was used to add the noise with the proper SNR [18]. This type of noise affects practically the whole spectrum, so one might expect less gain from a localized representation. As the results show (see Table III. again), the best 2D-DCT scores are indeed just slightly better than those obtained with MFCCs. One can also see, however, that in this case there was no significant difference between the performance of the spectrogram representation and the two best configurations of the critical-band-based method (53 channels and 26 channels).

VI. CONCLUSIONS

Our results support the findings of [10] in that the localized spectro-temporal features can result in the same or even better phone recognition accuracies than the conventional MFCC coefficients. We also showed that the same sort of 2D-DCT feature extraction can be performed on the conventional critical-band log-energy representation as well, yielding similar recognition accuracies while requiring less computational effort. Lastly, the experiments with noisy speech demonstrated that the proposed representation is less sensitive to additive noise than the conventional feature set. In the future we plan to run more experiments with various types of band-limited noise.

ACKNOWLEDGMENT

The authors would like to thank Jake Bouvrie for providing the pink-noise contaminated dataset.

REFERENCES

[1] X. Huang, A. Acero and H-W. Hon, Spoken Language Processing, *Prentice Hall*, 2001.

[2] T. Chih, P. Ru and S. Shamma, “Multiresolution spectrotemporal analysis of complex sounds,” *Journal of Acoust. Soc. America*, Vol. 118, pp. 887-906, 2005.

[3] M. Slaney, “An efficient implementation of the Patterson-Holdsworth auditory filter bank,” *Apple Computer Technical Report #35*, 1993.

[4] S. Furui, “On the role of spectral transition for speech perception,” *Journ. Acoust. Soc. America*, Vol. 80, No. 4, 1986, pp. 1016-1025.

[5] Young et al., The HTK Book (Manual of the Hidden Markov Model Toolkit HTK), *Cambridge University Engineering Department*.

[6] J.B. Allen, “How Do Humans Process and Recognize Speech?”, *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, pp. 567-577, 1994.

[7] H. Bourlard and N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, *Kluwer*, 1994.

[8] H. Hermansky and S. Sharma, “TRAPs – classifiers of temporal patterns,” *Proc. ICSLP'98*, pp. 1003-1006.

[9] M. Kleinschmidt, “Localized Spectro-Temporal Features for Automatic Speech Recognition, ” *Proc. EuroSpeech 2003*, pp. 2573-2576.

[10] J. Bouvrie, T. Ezzat, and T. Poggio, “Localized Spectro-Temporal Cepstral Analysis of Speech,” *Proc. ICASSP 2008*, pp. 4733-4736.

[11] L. Lamel, R. Kassel and S. Seneff, “Speech database development: Design and analysis of the acoustic-phonetic corpus,” *Proc. DARPA Speech Rec. Workshop*, pp. 100-109, 1986.

[12] K.-F. Lee and H. Hon, “Speaker-independent phone recognition using hidden Markov models,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1641–1648, Nov. 1989.

[13] C.M. Bishop, Neural Networks for Pattern Recognition, *Clarendon Press*, 1995.

[14] A. Halberstadt and J. Glass, “Heterogeneous measurements and multiple classifiers for speech recognition,” *Proc. ICSLP'98*, pp. 995-998.

[15] P. Clarkson and P.J. Moreno, “On the use of support vector machines for phonetic classification,” Proc. *ICASSP'99*, pp. 585-588.

[16] A. Kocsor, and L. Tóth, “Kernel-Based Feature Extraction with a Speech Technology Application,” *IEEE Trans. Signal Processing*, Vol. 52, No. 8, pp. 2250-2263, 2004.

[17] A. Varga, and H. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Communication, Vol. 12, No. 3, pp. 247-251, 1993.

[18] H-G. Hirsch: FaNT: Filtering and Noise-Adding Tool, http://dnt.kr.hs-niederrhein.de/download.html