LOCALIZED SPECTRO-TEMPORAL CEPSTRAL ANALYSIS OF SPEECH

Jake Bouvrie, Tony Ezzat, and Tomaso Poggio

Center for Biological and Computational Learning
Massachusetts Institute of Technology, Cambridge, MA
jvb@mit.edu, tonebone@mit.edu, tp@ai.mit.edu

ABSTRACT

Drawing on recent progress in auditory neuroscience, we present a novel speech feature analysis technique based on localized spectro-temporal cepstral analysis of speech. We proceed by extracting localized 2D patches from the spectrogram and project onto a 2D discrete cosine (2D-DCT) basis. For each time frame, a speech feature vector is then formed by concatenating low-order 2D-DCT coefficients from the set of corresponding patches. We argue that our framework has significant advantages over standard one-dimensional MFCC features. In particular, we find that our features are more robust to noise, and better capture temporal modulations important for recognizing plosive sounds. We evaluate the performance of the proposed features on a TIMIT classification task in clean, pink, and babble noise conditions, and show that our feature analysis outperforms traditional features based on MFCCs.

Index Terms— Speech processing, Speech recognition, Cepstral analysis, Nervous system

1. INTRODUCTION

Most state-of-the-art speech recognition systems today use some form of MEL-scale frequency cepstral coefficients (MFCCs) as their acoustic feature representation. MFCCs are computed in three major processing steps: first, a short-time Fourier transform (STFT) is computed from a time waveform. Then, over each spectral slice, a bank of triangular filters spaced according to the MEL-frequency scale is applied. Finally, a 1-D discrete cosine transform (1D-DCT) is applied to each filtered frame, and only the first N coefficients are kept. This process effectively retains only the smooth envelope profile from each spectral slice, reduces the dimensionality of each temporal frame, and decorrelates the features.

Although MFCCs have become a mainstay of ASR systems, machines still significantly under-perform humans in both noise-free and noisy conditions [13]. In the work presented here, we turn to recent studies of the mammalian auditory cortex [4, 16] in an attempt to bring machine performance towards that of humans via biologically-inspired feature analyses of speech. These neurophysiological studies reveal that cortical cells in the auditory pathway have two important properties which are distinctly *not* captured by standard MFCC features, and which we will explore in this work.

Firstly, rather than being tuned to purely spectral modulations, the receptive fields of cortical cells are instead tuned to both spectral and temporal modulations. In particular, auditory cells are tuned to modulations with long temporal extent, on the order of 50-200ms [4, 16]. In contrast, MFCC features are tuned only to spectral modulations: each 1D DCT basis may be viewed as a matched filter that responds strongly when the spectral slice it is applied to contains the spectral modulation encoded by the basis. MFCC coefficients thus indicate the degree to which certain spectral modulations are present in each spectral slice. The augmentation of MFCCs with Δ and $\Delta\Delta$ features clearly incorporates more temporal information, but this is not equivalent to building a feature set with explicit tuning to particular temporal modulations (or joint spectro-temporal modulations for that matter). Furthermore, the addition of Δ and $\Delta\Delta$ features creates a temporal extent of only 30-50ms, which is still far shorter than the duration of temporal sensitivities found in cortical cells.

Secondly, the above neurophysiological studies further show that cortical cells are tuned to *localized* spectro-temporal patterns: the spectral span of auditory cortical neurons is typically 1-2 octaves [4, 16]. In contrast, MFCC features have a *global* frequency span, in the sense that the spectral modulation "templates" being matched to the slice span the entire frequency range. One immediate disadvantage of the global nature of MFCCs is that it reduces noise-robustness: addition of noise in a small subband affects the entire representation.

Motivated by these findings, we propose a new speech feature representation which is localized in the time-frequency plane, and is explicitly tuned to spectro-temporal modulations: we extract small overlapping 2D spectro-temporal patches from the spectrogram, project those patches onto a 2D discrete cosine basis, and retain only the low-order 2D-DCT coefficients. The 2D-DCT basis forms a biologically-plausible matched filter set with the explicit joint spectro-temporal tuning we seek. Furthermore, by localizing the representation of the spectral envelope, we develop a feature set that is robust to additive noise.

2. BACKGROUND

A large number of researchers have recently explored novel speech feature representations in an effort to improve the performance of speech recognizers, but to the best of our knowledge none of these features have combined localization, sensitivity to spectro-temporal modulations, and low dimensionality.

Hermansky [7] and Bourlard [2] have used localized sub-band features for speech recognition, but their features were purely spectral and failed to capture temporal information. Subsequently, through their TRAP-TANDEM framework, Hermansky, Morgan and collaborators [7, 3] explored the use of long but thin temporal slices of critical-band energies for recognition, however these features lack joint spectro-temporal sensitivity. Kajarekar et al. [8] found that both spectral and temporal analyses performed in sequential order outperformed joint spectro-temporal features within a linear discriminant framework, however we have found joint 2D-DCT features to outperform combinations of purely spectral or temporal features. Atlas and Shamma [1] also explored temporal modulation sensitivity by computing a 1D-FFT of the critical band energies from a spectrogram. These features too lack *joint* and *localized* spectro-temporal modulation sensitivity. Kitamura et al. [9], take a *global* 2D-FFT of a MEL-scale spectrogram, and discard

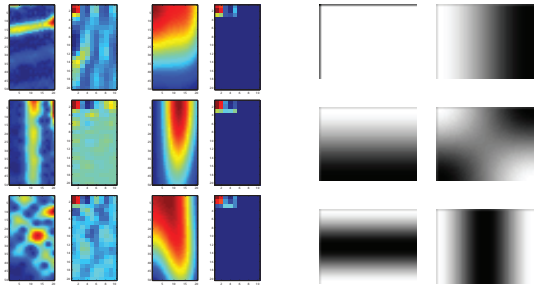

Fig. 1. Left two columns: Original spectrogram patches, followed by the corresponding 2D-DCT. Middle 2 columns: Patches reconstructed from low-order DCT coefficients followed by the low-order DCT coefficients retained for reconstruction. Last 2 columns: Retained low-order 2D-DCT basis functions.

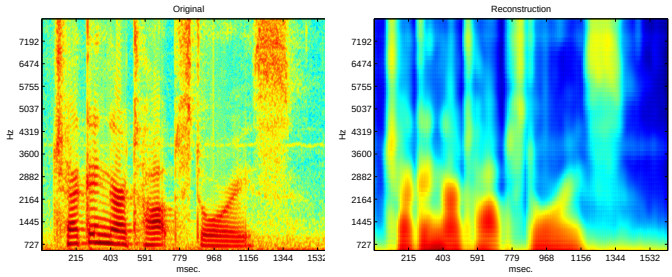

Fig. 2. Left: Original spectrogram. Right: Resulting spectrogram after retaining only low-order DCT coefficients per patch, and applying overlap-add reconstruction.

various low-frequency bands from the resulting magnitude. This approach does not provide any joint spectro-temporal localization, and cannot be interpreted as capturing meaningful configurations of specific spectro-temporal modulation patterns. It is the localized analysis in our method, and the fact that we seek to encode spatial configurations of important spectro-temporal modulations, that critically differentiates our approach from much of the previous work.

Perhaps the closest work to ours is that of Shamma and colleagues [4], and the work of Kleinschmidt, Gelbart, and collaborators [10, 11]. In [4] the authors apply localized complex filters that produce both magnitude and phase information for the purpose of speech vs. non-speech detection. The latter group applied data-optimized Gabor filters to blocks of MEL-scale spectra with 23 frequency bins, and then present ASR results when Gabor features augment other features. Our work builds on upon both of these efforts, and demonstrates an important point which we believe has not been made strongly enough in these previous works: that a simple set of localized 2D-DCT features (in this case, "bar-like" detectors faithful to the auditory neuroscience) is on its own powerful enough to achieve state-of-the-art performance on a difficult phonetic discrimination task.

3. 2-D CEPSTRAL ANALYSIS OF SPEECH

3.1. Spectro-Temporal Patch Extraction

The first step of our technique is to compute the log-magnitude of the STFT of the signal. We then normalize the resulting (log) spectrogram to have zero mean and unit variance. Then, at every grid point (i, j) in the spectrogram, we extract a patch $P_{ij}(f, t)$ of size df and width dt. The height df and width dt of the local patch are important analysis parameters: they must be large enough to be able to resolve the underlying spectro-temporal components in the patch, but small enough so that the underlying signal is stationary. Additional analysis parameters are the 2D window hop-sizes in time Δi and frequency Δj, which control the degree of overlap between neighboring patches. Finally, we pre-multiply the patch with a 2D Hamming window $W_H(f, t)$ in order to reduce border effects during subsequent patch processing.

3.2. 2D Discrete Cosine Transform and Coefficient Truncation

After patch extraction, a 2-D discrete cosine transform (2D-DCT) is applied to each windowed patch $P(f, t)$ to produce a set of DCT coefficients $B(\Omega, \omega)$. The 2D-DCT projects each patch onto a set of orthogonal, separable cosine basis functions that respond to "horizontal" speech phenomena such as harmonics and formants, "vertical" speech phenomena such as plosive edges, and more complex spectro-temporal noise patterns. In the rightmost two columns of Figure 1 we show the six low-order 2D-DCT basis functions used in our analysis. The top-left basis is everywhere uniform. Shown in Figure 1 in the first column are representative harmonic (top), plosive (middle), and noise (bottom) patches from a spectrogram, along with their respective 2D-DCT coefficients in the second column. As expected, horizontal harmonic edges in a patch strongly activate coefficients in the first column of the corresponding DCT, and vertical plosive phenomena activate DCT coefficients in the first row. Noise phenomena, with more high frequency components than the previous two examples, has energy that is distributed among most of the DCT coefficients.

The last step of our analysis consists of truncating the 2D-DCT and retaining only the low-order coefficients for each patch. The effect of doing this is also shown in Figure 1: original patches in the first column are reconstructed in the third column using only the low-order 3×5 block of DCT coefficients (4th column). Keeping only the low-order DCT coefficients is equivalent to representing each patch with a *smooth spectro-temporal envelope*. We further illustrate this concept in Figure 2, where the original spectrogram displayed on the left is reconstructed on the right from low-order patch 2D-DCT coefficients. The individually reconstructed patches are overlap-added together to assemble the full spectrogram. In this example, we have used analysis windows of size 780Hz by 57ms shifted in steps of 156Hz in frequency and 10ms in time.

4. CLEAN SPEECH TIMIT EXPERIMENTS

The above 2D-DCT analysis was applied towards extracting features for phonetic classification on the TIMIT corpus [12], and compared to MFCC-based features used by Clarkson and Moreno [5], and the best single set of features proposed by Halberstadt and Glass in [6]. The latter feature set is the best baseline that we are aware of. We divided TIMIT into standard train and test sets following the convention in [15, 6]. The 2D-DCT analysis is performed using both wideband and narrowband spectrograms for comparison. After ignoring glottal stops ('q' tokens), the 60 remaining phonetic classes are later mapped to 39 categories after training but before scoring, also following standard practices.

4.1. Spectrogram Pre-processing

TIMIT utterances are first normalized and subjected to a pre-emphasis filter. We then compute spectrograms using 32 sample (2ms) hops, 1024-point FFTs and 300 sample (18.75ms) Hamming windows or 150 sample (9.375ms) windows for narrow- and wideband conditions respectively. We then take the log-magnitude of the

resulting spectrum and normalize to give a global utterance mean of zero and unit variance. Utterances are broken up in time according to the labeled phonetic boundaries and enlarged by an additional 30ms on either side of each phoneme so as to include coarticulatory phenomena. Each resulting independent phoneme is then truncated at 6.23kHz (400 frequency bins), while a copy of the bottom 25 low-frequency bins is reflected, for all time, about the 0Hz edge and appended. Because we later apply local 2D-DCTs to Hamming windowed regions of the spectrogram, reflection is done to avoid artificially down-weighting low frequency bins near the edge of the image.

4.2. 2D-DCT Patch Processing and Coefficient Truncation

We first compute a sliding localized two-dimensional DCT over the phoneme's spectrogram. While many reasonable window and step sizes exist, we have found that, for the narrowband STFT parameters above, good results are obtained with 780Hz by 56.75ms (or 50 by 20 bin) Hamming windowed 2D analysis regions with a 390Hz (25-bin) frequency step-size and a 4ms (2-bin) time step-size. For wideband STFT conditions, good results are obtained with 623Hz by 107.375ms (or 40 by 50 bin) Hamming windowed 2D analysis regions with identical step sizes in time and frequency as in the narrowband case. The 2D-DCT is computed with 2x oversampling in both time and frequency. We have found that performance does not critically depend on the precise window and step size choices above. To avoid implicit overfitting, evaluation of performance for different parameter choices was done using the TIMIT development set proposed by [6], while the final evaluations shown below were done on the core test set.

For each 2D analysis region, we save only the 6 lowest-order 2D-DCT coefficients corresponding to the upper left 3×3 triangle in the DCT image. These coefficients collectively encode only the patch's DC offset, two low spatial frequency horizonal and vertical basis components, and one "checkerboard" basis component. Saving six coefficients per patch at the above resolution smooths out any remaining harmonic structure, leaving only the spectro-temporal envelope.

4.3. Feature Vector Construction

The previous step provides a vector of 6 features for each patch taken from the phoneme. We modify the approach of [6] in order to compute a fixed length feature vector from the variable number of 2D-DCT coefficients representing a given phoneme; this particular construction was found (in [6]) to work well for the MFCC-based features computed therein. If the 6-dimensional vectors are collected and arranged in a relative order corresponding to the time-frequency centers of the respective analysis windows, we are left with a 3D matrix of coefficients per phoneme example: $S(i, j, k)$ where i indexes time, j indexes frequency, and k is the DCT coefficient index. The number of time bins will of course vary across phonemes. We therefore divide up the time axis of the 3D matrix of coefficients into five segments, and average over time within each segment. The time bins corresponding to the 30ms of additional signal added before and after the phoneme give the first and last segments, while the bins falling within the phoneme itself are divided up in 3:4:3 proportion to give the middle 3 segments. All coefficients across the five averaged segments (contributing 17 patches $\times$ 6 coefficients $= 102$ features each) are then pooled and concatenated into a single 510-dimensional vector. Lastly, the log-duration of the phoneme is added to give the final 511-element feature vector. Prior to classification, the training and test datasets are whitened with the principal components derived from the training set.

Features	Stops	Vowels	All	Dims
CM	29.66	37.59	28.30	196
HA	27.91	37.80	25.60	61
2D-DCT-NB	23.53	37.33	24.93	511
2D-DCT-WB	25.53	36.69	24.37	511
2D-DCT-NB/SVM2			21.37	

Table 1. Percent error rates for the three sets of features when training/testing on stops only, vowels only, or on all phonemes from clean utterances. Our features are denoted "2D-DCT". Dimensionality of the feature vectors are given in the last column. "NB" and "WB" denote narrowband and wideband conditions respectively.

4.4. MFCC-Based Baseline Comparison Features

We compare our features to two other TIMIT MFCC-based baseline feature sets: the "S2" feature-set proposed by Halberstadt & Glass [6] and the features described by Clarkson & Moreno [5]. We will refer to these feature sets as "HA" and "CM" respectively. The HA feature set is constructed by computing 12 MFCCs from each frame of the spectrogram. Temporal averages are taken over the five non-overlapping segments described above to obtain a fixed-length feature vector, and a log-duration feature is added. For CM features, 13 MFCCs are computed for each spectrogram frame. However, Δ and $\Delta\Delta$ features are also included, giving classical 39-dimensional feature vectors for each frame. The time axis is again divided up into five segments, but the two regions including spectra before and after the phoneme are 40ms wide and are centered at the beginning and end of the phoneme. A log-duration feature is also added. In HA and CM, the resulting datasets are whitened with PCA.

4.5. Classification Framework

All-vs-all (AVA) classification with linear regularized least-squares (RLS) classifiers [14] was performed on the resulting datasets. We include for comparison results on the full TIMIT task using second-order polynomial SVMs with 5-fold cross-validated selection of the regularization parameter. Our ultimate goal, however, is to illustrate the strength of localized spectro-temporal features even in the absence of excessive tuning of the classifier stage.

4.6. Clean Speech Results

In Table 1 we show linear RLS classification error rates for the proposed localized 2D-DCT features (for both narrow- "NB" and wideband "WB" spectrograms) as compared to the two sets of baseline features described above. We show results on the full TIMIT task, and additionally, when training and testing on subsets consisting of just the vowels or just the stops. The full task consists of training on 60 classes (140225 train, 7215 test examples) and then mapping to 39 classes for scoring, while the stops task consists of 6 phonetic classes (16134 train, 799 test examples) and the vowel task consists of 20 classes (45572 train, 2341 test examples). No post-mapping is done prior to scoring in the case of the the vowels and stops experiments.

In all cases, the localized 2D-DCT features outperform the MFCC-based baseline features. Wideband spectrograms with longer temporal analysis extents are seen to give better performance than narrowband spectrograms with shorter extents in all experiments excepting the stops only evaluation. However in the case of stops in particular, the 2D-DCT features provide substantial improvement over traditional MFCCs. Because the DCT analysis is spectro-temporal and includes explicit bases encoding vertical and horizontal spatial gratings, the 2D-DCT features capture the strong vertical "edges" present in stops and other plosives. The last row of Table 1

Features	Clean	20dB	10dB	0dB
CM-RLS1	28.30	61.68	79.67	91.78
HA-RLS1	25.60	41.34	63.12	80.03
2D-DCT-NB/RLS1	24.93	32.53	48.16	71.93
2D-DCT-WB/RLS1	24.37	32.36	47.79	72.75
2D-DCT-NB/RLS1 (B)	24.93	38.99	59.76	77.28
2D-DCT-WB/RLS1 (B)	24.37	37.73	57.30	75.05

Table 2. Train clean, test noisy experiments: Percent error rates on the full TIMIT test set for several signal-to-noise ratios and feature sets.

shows performance when using nonlinear SVM classifiers, and confirms that 2D-DCT features still exhibit the reduction in error that one would expect when moving to more complex classifiers.

5. NOISY SPEECH TIMIT EXPERIMENTS

The classification performance of localized 2D-DCT features was also evaluated in the presence of both pink and babble noise. In the case of pink-noise, we provide a comparison with HA [6] and CM [5] features. The HA-RLS1 error rates were originally presented in Rifkin et al. [15], and are reproduced here. The authors of [15], do not provide performance in babble-noise. In all experiments, *training is done on clean speech while testing is done on noisy speech*.

5.1. Noisy Dataset Construction

Pink noise corrupted TIMIT utterances at 20dB,10dB, and 0dB SNR were obtained from the authors of [15] so that experiments could be performed under the exact same noise mixing conditions. In [15], a single 235 second segment of noise from the NOISEX-92 dataset was used to artificially corrupt the test set speech. Random contiguous snippets from this master segment were added to each utterance with an amplitude chosen to satisfy the desired global SNR. Similarly, we constructed babble-noise corrupted TIMIT utterances by following the same procedure while using a 235 second segment of babble-noise, also from the NOISEX-92 dataset. In both cases, spectra and features were then extracted from the noisy utterances in a manner identical to that described in Section 4.

5.2. Noisy Speech Results

In Table 2 we show percent error rates for the full TIMIT phonetic classification task, comparing the HA and CM feature sets to the proposed localized 2D-DCT features when using linear RLS classifiers with an all-vs-all multiclass scheme (denoted "RLS1"). The first four feature set/classifier combinations involve pink-noise corrupted utterances.

In the presence of even weak pink noise (e.g. 20dB), 2D-DCT features with simple linear classifiers outperform HA features. As the signal to noise ratio is decreased, the performance advantage remains significant: we observe a relative reduction in error of approximately 10-25% when using localized 2D-DCT features over the MFCC-based HA features with an identical classification stage. Despite the fact that the CM feature set combines MFCCs with traditional Δ and $\Delta\Delta$ features, both the DCT and HA features far outperform Clarkson's CM features. In the last two rows of Table 2 (marked with a "B"), we show classification error in babble noise. Although babble noise is usually considered a more challenging condition for speech recognizers, for this particular task we observe only a modest increase in error above the error in pink-noise when using the proposed 2D-DCT features. The longer temporal extents of the patches and 2D-DCT templates in the "WB" case are also seen to give improved performance in babble noise.

6. DISCUSSION

The biologically inspired feature analysis presented in this paper consisted of two main steps: (1) Extraction of localized spectro-temporal patches, and (2) low-dimensional spectro-temporal tuning using the 2D-DCT. A localized encoding and extraction of structure in the time-frequency plane faithfully preserves the general distribution of energy, and retains critical discriminatory information. In Table 1 we showed that discrimination among phonemes with strong temporal modulation, such as plosives and stop consonants, was better with local 2D-DCT features than with the two sets of "global" MFCC-based baseline features. Traditional cepstral analysis, while admittedly lower dimensional in many cases than the proposed features, tends to over-smooth in frequency and ignore important dynamics of the spectro-temporal envelope.

In pink-noise corrupted speech, local 2D-DCT features provide substantial additional noise-robustness beyond the baseline MFCC features as measured by classification accuracy on the TIMIT corpus. We also found that the localized 2D-DCT analysis outperforms classical MFCCs augmented with delta and acceleration features. Although this feature set is not the strongest of the two baselines, the comparison shows that even features incorporating more temporal information per frame is not sufficient; both time and frequency localization is necessary. On the whole, the phoneme classification experiments presented above show that the method is viable, outperforming a state-of-the-art baseline in clean and noisy conditions.

7. REFERENCES

[1] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[2] H. Bourlard and S. Dupont, "Subband-based speech recognition," in *Proc. ICASSP 97*, 1997.

[3] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in mvcsr using neural networks," in *Proc. ICSLP*, 2004.

[4] T. Chih, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 2005.

[5] P. Clarkson and P.J. Moreno. "On the use of support vector machines for phonetic classification," in *Proc. IEEE ICASSP*, Vol 2, pp.585–588, 1999.

[6] A. Halberstadt and J. Glass. "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proc. ICSLP*, 1998.

[7] H. Hermansky. "Trap-tandem: Data-driven extraction of temporal features from speech,"in *Proc. ASRU Workshop*, 2003.

[8] S. Kajarekar, B. Yegnanarayana and H. Hermansky. "A Study of Two Dimensional Linear Discriminants for ASR", in *Proc ICASSP*, Salt Lake City, Utah, USA, May, 2001.

[9] T. Kitamura, E. Hayahara, and Y. Simazaki, "Speaker-independent word recognition in noisy environments using dynamic and averaged spectral features based on a two-dimensional mel-cepstrum," in *Proc. ICSLP*, 1990.

[10] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. Eurospeech*, 2003.

[11] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *Proc. ICSLP*, 2002.

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

[13] R. P. Lippmann. "Speech recognition by machines and humans", *Speech Communication*, 22(1), pp.1-15, 1997.

[14] R. Rifkin. "Everything Old Is New Again: A Fresh Look at Historical Approaches to Machine Learning", Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[15] R. Rifkin, K. Schutte, D. Saad, J. Bouvrie, and J. Glass. "Noise Robust Phonetic Classification with Linear Regularized Least Squares and Second-Order Features", in *Proc. IEEE ICASSP*, 2007.

[16] F.E. Theunissen, K. Sen, and A. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *J. Neuro.*, Vol. 20, pp.2315–2331, 2000.