Spectro-Temporal Directional Derivative Features for Automatic Speech Recognition

James Gibson, Maarten Van Segbroeck, Antonio Ortega, Panayiotis Georgiou, and Shrikanth Narayanan

Signal and Image Processing Institute, University of Southern California, Los Angeles, CA

jjgibson@usc.edu, {maarten, ortega, georgiou, shri} @sipi.usc.edu

Abstract

We introduce a novel spectro-temporal representation of speech by applying directional derivative filters to the Mel-spectrogram, with the aim of improving the robustness of automatic speech recognition. Previous studies have shown that two-dimensional wavelet functions, when tuned to appropriate spectral scales and temporal rates, are able to accurately capture the acoustic modulations of speech, even in high noise conditions. Therefore, spectro-temporal features extracted from the wavelet transformation of the spectrogram, offer additional noise robustness to important signal processing tasks, such as voice activity detection and speech recognition. In this paper, we explore the use of the steerable pyramid, a directional wavelet transform that is common in image processing, to derive a spectro-temporal feature representation of speech that can serve as an alternative to cepstral derivatives and Gabor filter-bank features. We discuss their application for the task of robust automatic speech recognition. Experiments conducted on the Aurora-2 database demonstrate their competitive robustness to other state-of-the-art speech features, especially in low signal-to-noise ratio conditions.

Index Terms: spectro-temporal features, automatic speech recognition, directional wavelet transforms

1. Introduction

Directional wavelets are popular for multi-resolution analysis of image and video data. In the last decade, Kleinschmidt and others have proposed using wavelet basis functions for multi-resolution, spectro-temporal representations of the speech spectrogram image [1, 2]. The Gabor wavelet is the most common of these directional filter-bank representations of speech due to its similarity to the auditory perception systems of mammals [3]. Other biologically inspired spectro-temporal speech features, e.g., cortical features, have been proposed and shown to be successful in a variety of speech related tasks including assessing speech intelligibility [4] and voice activity detection [5]. In this paper, we focus on Gabor filter-bank (GFB) features as our baseline since they are not only similar to those proposed here, but have shown success in a variety of speech processing applications.

Gabor filter-bank speech features have been shown to be noise robust in applications including ASR and voice activity detection (VAD) [6–9]. One of the merits of Gabor features is their ability to more readily incorporate longer term information than offered by Mel-frequency cepstral coefficients (MFCCs), which are typically taken over a very short frame (approximately 25 ms); long-term temporal information with MFCCs is simply incorporated by computing first and second order derivatives to yield the Delta (Δ) and Delta-Delta ($\Delta\Delta$) features in a compact manner. While they have been widely adopted due to their ability to model onset and offset regions of the spectral energy fluctuations, they do not capture spectro-temporal modulations in speech. Gabor features allow for that possibility. Unfortunately, this is at the cost of having very large dimensional features. Dimensionality reduction methods, e.g., principal component analysis, have been applied to Gabor features [10]. However, these methods require training data and additional computational overhead.

In this paper, we introduce Directional Derivative (DD) features that provide the relative merits of each of these well studied representations. These new features exhibit both the noise robustness found in Gabor features and benefit from dimensionality reduction by the discrete cosine transform (DCT) as in MFCC features. We apply the 2D Directional Derivative transformation to the speech spectrogram image and demonstrate the efficacy of the resulting representation on a continuous digit speech recognition task with the Aurora-2 corpus.

2. Methodology

We use the log Mel-spectrogram with 23 Mel-bands as the time-freqency representation from which all subsequent spectro-temporal features are computed. The log Mel-spectrogram is computed using 25 ms windows with a 10 ms window shift. It should be noted that the general method proposed here can be applied to other speech time-frequency representations such as the Gammatone spectrogram [11], the modulation spectrogram [12], and the auditory spectrogram [13], however, this remains a topic for future investigation.

2.1. Directional Derivative Features

Directional Derivative features are dynamic features that incorporate long-term signal variability as well as spectro-temporal modulations via directional filters. We choose a wavelet transform for this task because it readily allows for multi-resolution, spectro-temporal analysis without introducing largely redundant features (due to critical or near-critical sampling). Critical sampling helps to drastically reduce feature dimensions for Gabor features [8,10]. However, it is important to note that the critical sampling is performed after spectro-temporal filtering thus requiring a much higher number of computations. We propose using a transform that offers a spectro-temporal analysis with an efficient implementation and low dimensional representation.

2.1.1. *The Steerable Pyramid Wavelet Transform*

The steerable pyramid wavelet was introduced by Simoncelli and others for image coding [14, 15]. It is a translation and rotation invariant wavelet representation with directional derivative bases. The directional derivative filters are oriented with respect to the X-Y plane, in this case frequency bins vs time frames, where a 90° orientation differentiates with respect to the temporal axis (like Δ features) and a 0° orientation differentiates with respect to the spectral axis and orientations between capture spectro-temporal modulations.

The steerable pyramid filter-bank operates on the image by first separating the image into a 2D high pass and 2D low pass channel. Next, the low pass channel is down-sampled by a factor of two with respect to both dimensions. The down-sampled low pass images are then passed through the directional filters. This process is then repeated on each subsequent low pass channel to offer a multi-resolution representation. The output of each directional sub-band is concatenated to the others to create the feature vectors. When applying this transform to a spectrogram image it is necessary to resample the sub-band images with respect to the time axis to produce a feature vector for every time frame and to have consistent duration with the original spectrogram image.

For this work we chose 90°, 62.5°, 45°, -45°, and -62.5° orientations for the directional filters. We did not include any angles between -45° and 45° (e.g., -22.5°, 0°, and 22.5°) because we observed these angles do not capture patterns relevant to the speech recognition task (i.e., a derivative with respect to frequency is not localized in time and therefore presents no phoneme specific information). The optimal number, and which particular, orientations would give the best results for various speech modeling tasks is still an open problem. Existing literature investigating this for other spectro-temporal features suggest that the optimal orientations are likely task, and possibly data, specific [10]. We will investigate this issue in more depth in subsequent studies. We will also investigate higher order directional derivative filters.

Figure 1(a) shows the impulse response of the directional sub-bands of the steerable pyramid transform used for this work. We do not include the high pass residue in the final feature representation as it is very sensitive to noise. For the proposed steerable pyramid implementation we use Simoncelli's steerable pyramid toolbox [16].

2.1.2. *Sub-band Dimensionality Reduction*

The discrete cosine transform is used to reduce feature dimensions when certain assumptions can be made about the signal (or image) on which it is being applied. The assumption is that the input has most of its energy confined to certain coefficients of the transform, typically the lower coefficients, due to an assumption that the signal is smooth. DCT dimensionality reduction is the final stage in MFCC computation and it both provides a lower dimensionality than Mel-frequency spectral features and offers de-correlated feature dimensions.

Because of the nature of the filters employed in the steerable pyramid transform, it is appropriate to assume a smooth representation with respect to the spectral axis. However, due to the directionality of the filters the number of coefficients required to retain most of the energy varies between sub-bands. Figure 1(b) shows the impulse response of the filter-bank sub-bands after the DCT is applied with respect to the spectral axis. The middle sub-band, corresponding to the filter with 90° orientation, has the majority of its energy concentrated in the first

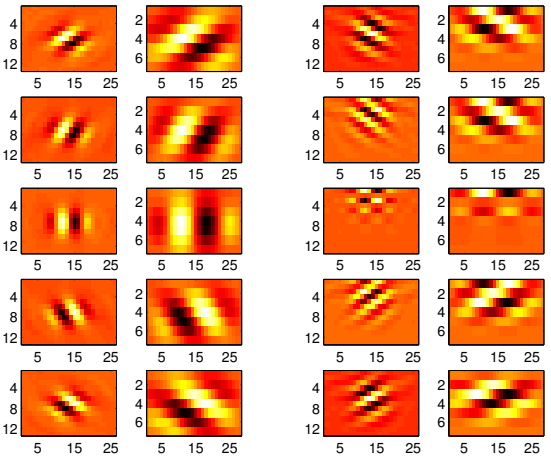

Figure 1: *Impulse response of the steerable pyramid filter-bank* (a) *before and* (b) *after the DCT is applied.*

half of its spectral coefficients. This is intuitive because this filter is akin to velocity delta features for MFCCs where the first half of the DCT coefficients are retained. For the other orientations the number of coefficients required to retain the majority of the energy, grows as the filter becomes less aligned with the vertical direction (i.e., 45° requires more coefficients than 62.5°, which in turn requires more than 90°). We use an affine function to determine the number of coefficients to retain that is proportional to the orientation degree, (θ). Specifically we compute the number of coefficients to keep from each sub-band according to:

$$\#\text{ coefficients} = \left\lceil \left(1 - \frac{|\theta^\circ|}{180^\circ}\right) N \right\rceil + 1, \qquad (1)$$

where N is the number of spectral bins in the sub-band. This is an empirical choice reached through inspection of the individual sub-bands to ensure that the majority of energy in each sub-band is retained. Figure 2 shows the coefficient energy distribution of the directional sub-bands after the DCT is applied (white indicates high energy, black low). Clearly, the energy content is dependent upon the orientation angle of the directional filter. The blue line indicates the number of coefficients retained according to eq. 1.

2.2. Comparison with baseline features

We show a comparison of feature dimensions in Table 1. MFCC-$\Delta\Delta$ is the traditional static, velocity, and acceleration feature representation. For the remainder of this paper we use the notation $\Delta\Delta$ to indicate that both velocity and acceleration features are included. GFB2 and GFB3 correspond to two levels (0 and 6.2 Hz temporal modulation filters) and three levels (0, 6.2, and 9.9 Hz temporal modulation filters), respectively, of the Gabor filter-bank being applied to the spectrogram image, where the first level is comprised of static (0 Hz) 2D low pass filters. MFCC-DD1 and MFCC-DD2 correspond to MFCCs augmented by the Directional Derivative features with one level and two levels of decomposition. As this table shows MFCCs

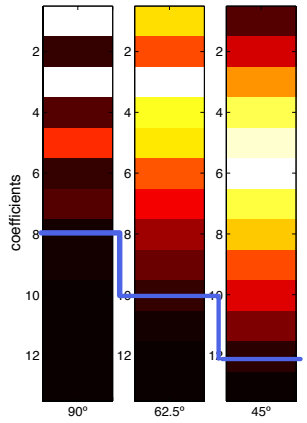

Figure 2: *Impulse response coefficient energy after the DCT of the directional sub-bands.*

augmented by Directional Derivative features still retain a low dimensionality relative to Gabor features while still offering spectro-temporal modeling via the directional filter-bank.

	Directional	Dimensions
MFCC-ΔΔ	No	39
GFB2	Yes	104
GFB3	Yes	173
MFCC-DD1	Yes	58
MFCC-DD2	Yes	82

Table 1: *Comparison of feature properties.*

Figure 3 shows how each of the contextual features are computed. It is important to note that downsampling is done before 2D convolution followed by linear interpolation resampling, with respect of the temporal axis, to produce the same number of feature vectors as the static spectrogram image. Downsampling first rather than critically sampling the sub-bands after the filter-bank gives a more efficient implementation as fewer computations are required on the down-sampled image.

3. Experiments and Results

We use the Aurora-2 corpus to examine the efficacy of the proposed features with traditional MFCC-ΔΔ and Gabor features (code implemented by Schädler et al. [17]) serving as baselines [18]. The Aurora-2 corpus has a training set comprised of 8,440 utterances and three test sets each comprised of 1,001 utterances. It has eight noise types (which are listed in table 3). All the experiments in this work are trained using clean utterances from the training set. We use HTK [19] for recognition, with whole word left-to-right Hidden Markov Models (HMMs) with 16 states per word and twenty diagonal covariance Gaussians per state. We use mean-variance normalization for all the compared features as this has been shown to offer additional noise robustness [8].

We show the word recognition accuracy (WRA) (averaged over all test sets and noise types) of the competing schemes in table 2. The highest WRA for each SNR is shown in **bold**. All the tested schemes perform similarly with respect to the higher SNRs (15 and 20 dB). However, at lower SNRs the performance

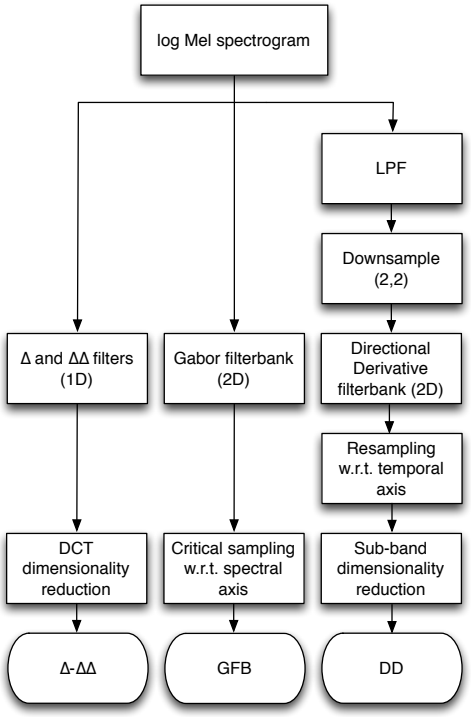

Figure 3: *Comparison of contextual feature extraction methods for one level of decomposition*

of MFCC-ΔΔ features suffer from the effects of noise significantly ($p < 0.005$ at 0 dB using the Wilcoxon signed-rank test) more than the other baseline scheme (GFB2). We find that adding an additional filter-bank level (9.9 Hz) for Gabor features degrades performance at low SNRs. The relative importance of particular modulation filters has been explored in more depth in [20] and [10]. We use all nine spectral modulation filters (0, ±0.03, ±0.06, ±0.12, and ±0.25 cyc./ch.) in the GBF feature extraction methods to give a direct comparison to previous literature.

	20dB	15dB	10dB	5dB	0dB
MFCC-ΔΔ	98.4	96.4	91.0	75.6	48.0
GFB2	98.6	97.2	92.7	81.6	59.1
GFB3	**99.0**	**97.5**	92.7	79.1	50.5
MFCC-GFB2	98.7	97.1	92.6	81.2	57.2
MFCC-DD1	98.5	97.0	92.6	80.5	56.0
MFCC-DD2	98.7	97.1	**93.0**	**82.3**	**60.6**

Table 2: *Word Recognition Accuracy (%) averaged across all test sets and noise types in the Aurora 2 corpus.*

We use GFB2 features concatenated with static MFCCs as an additional baseline, which allows for more direct comparison as the Directional Derivative features serve as dynamic features to augment static MFCCs. MFCCs augmented by Gabor features perform marginally worse than GFB2 alone. This is likely because the first level of the Gabor filter-bank is a lowpass channel that captures similar information as static MFCCs resulting in redundant features by concatenating the two. Schädler et al. found that using multi-layer perceptron modeling of the concatenated MFCC and GFB features results in an improvement

over using them independently [10], however we do not explore any TANDEM-based approaches (e.g., [21]) in this work.

Our proposed features result in the highest WRA for 0, 5, and 10 dB SNR (The difference is significant with $p < 0.05$ for 0 and 5 dB). GFB3 features result in the highest WRA for both 20 and 15 dB SNR. However, neither of these differences are significantly higher than the GFB2 or Directional Derivative features.

Table 3 shows how the Directional Derivative features compare with the 2-level Gabor features for each noise type. Our proposed features perform better for all noise types except for additive Subway noise (0.3% absolute reduction) and Train-Station noise (1.1% absolute reduction).

Test Set	Noise Type	GFB2	MFCC-DD2
A	Subway	**86.9**	86.6
	Babble	84.2	**85.4**
	Car	86.3	**86.7**
	Exhibition	82.3	**83.9**
B	Restaurant	85.9	**87.1**
	Street	85.9	**86.4**
	Airport	87.1	**87.9**
	Train-Station	**87.5**	86.4
C	Subway	86.4	**86.5**
	Street	85.8	**86.3**

Table 3: *Word Recognition Accuracy (%) averaged across 0-20 dB SNRs.*

4. Conclusions and Future Work

We presented a novel spectro-temporal wavelet feature that is very competitive with both MFCC-$\Delta\Delta$ features and spectro-temporal Gabor features, especially at low SNRs. We compared performance of these schemes for a continuous digit recognition task. In the future, we plan to extend this work to alternative time-frequency representations such as the Gammatone spectrogram and the auditory spectrogram. We also plan to explore the efficacy of other common directional wavelet transforms, such as Curvelets [23] and Contourlets [24], for the speech modeling task. Furthermore, we are interested in comparing our proposed framework other methods of incorporating contextual information for speech recognition such as Linear Discriminant Analysis (LDA) and the Karhunen-Loeve Transform (KLT) on stacked short-time feature vectors [25]. Finally, we want to investigate how the benefits of the proposed methods translate into large-vocabulary continuous speech recognition (LVCSR) tasks especially when combined with a range of other performance boosting techniques (e.g. discriminative training).

5. Acknowledgements

This work was funded in part by the USC Annenberg Fellowship Program.

6. References

[1] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *Proc. ICSLP*, vol. 5, 2002, pp. 16–38.

[2] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. Eurospeech*, vol. 87. Citeseer, 2003.

[3] N. Mesgarani, S. David, and S. Shamma, "Representation of phonemes in primary auditory cortex: how the brain analyzes speech," in *Proc. ICASSP*, vol. 4, 2007, pp. IV–765.

[4] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech communication*, vol. 41, no. 2, pp. 331–348, 2003.

[5] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 920–930, 2006.

[6] B. Meyer, S. Ravuri, M. Schädler, and N. Morgan, "Comparing different flavors of spectro-temporal features for asr," in *Proc. of Interspeech*, 2011, pp. 1269–1272.

[7] B. Meyer, C. Spille, B. Kollmeier, and N. Morgan, "Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition," in *Proc. Interspeech*, vol. 15, 2012, p. 20.

[8] M. R. Schädler and B. Kollmeier, "Normalization of spectro-temporal gabor filter bank features for improved robust automatic speech recognition systems," in *Proc. Interspeech*, 2012.

[9] T. Tsai and N. Morgan, "Longer features: They do a speech detector good," in *Proc. Interspeech*, 2012.

[10] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, p. 4134, 2012.

[11] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *APU report*, vol. 2341, 1988.

[12] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1, pp. 117–132, 1998.

[13] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 824–839, 1992.

[14] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 587–607, 1992.

[15] E. Simoncelli and W. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proc. ICIP*, vol. 3, 1995, pp. 444–447.

[16] E. Simoncelli. (2003) Steerable pyramid toolbox. [Online]. Available: http://www.cis.upenn.edu/ eero/steerpyr.html

[17] M. R. Schädler. (2011) Gabor filter bank (gbfb) feature extraction reference implementation in matlab. [Online]. Available: http://medi.uni-oldenburg.de/56428.html

[18] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[19] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The htk book," *Cambridge University Engineering Department*, vol. 3, 2002.

[20] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," in *Proc. Eurospeech*, vol. 97, 1997, pp. 1079–1082.

[21] H. Hermansky and P. Fousek, "Multi-resolution rasta filtering for tandem-based asr," in *Proc. Interspeech*, 2005.

[22] S. Thomas, S. Ganapathy, and H. Hermansky, "Tandem representations of spectral envelope and modulation frequency features for asr," in *Proc. Interspeech*, 2009.

[23] E. Candes and D. Donoho, "Curvelets, multiresolution representation, and scaling laws," in *Proc. SPIE*, vol. 4119, no. 1, 2000.

[24] M. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *Image Processing, IEEE Transactions on*, vol. 14, no. 12, pp. 2091–2106, 2005.

[25] B. Milner, "Inclusion of temporal information into features for speech recognition," in *Proc. ICSLP*, vol. 1, 1996, pp. 256–259.