Comparing Time-Frequency Representations for Directional Derivative Features

James Gibson, Maarten Van Segbroeck, and Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA

http://sail.usc.edu

Abstract

We compare the performance of Directional Derivatives features for automatic speech recognition when extracted from different time-frequency representations. Specifically, we use the short-time Fourier transform, Mel-frequency, and Gammatone spectrograms as a base from which we extract spectro-temporal modulations. We then assess the noise robustness of each representation with varied number of frequency bins and dynamic range compression schemes for both word and phone recognition. We find that the choice of dynamic range compression approach has the most significant impact on recognition performance. Whereas, the performance differences between perceptually motivated filter-banks are minimal in the proposed framework. Furthermore, this work presents significant gains in speech recognition accuracy for low SNRs over MFCCs, GFCCs, and Directional Derivatives extracted from the log-Mel spectrogram.

Index Terms: time-frequency representations, spectro-temporal features, automatic speech recognition, noise robustness

1. Introduction

Spectro-temporal modulation features offer increased noise robustness for several speech applications including automatic speech recognition (ASR) [1–3]. The typical scenario for extracting spectro-temporal modulations from speech is to filter an appropriate time-frequency representation with two dimensional derivative-like filters oriented in the time-frequency plane. The choice of time-frequency representation is essential for providing relevant speech information to the subsequent spectro-temporal filtering stage.

Several time-frequency representations have been introduced for speech processing. Many of them make use of human perception or biology such as the Mel-frequency spectrogram, perceptual linear prediction (PLP) [4], the cocleagram (Gammatone spectrogram) [5], and the auditory spectrogram [6]. To further make use of human inspired auditory knowledge, there are many post processing steps that have been applied to these time-frequency representations such as: the discrete cosine transform (DCT), which provides energy compaction; RASTA filtering, which models humans' insensitivity to slow varying stimuli [7]; and Gabor features, which use spectro-temporal modulation filters to approximate processing done by the auditory system [8–10].

In [11], the authors propose Gabor features extracted from the Power Normalized Spectrogram, a version of the Gammatone spectrogram with power-law compression and power bias subtraction. They reported a large performance gain by using this representation over previous Gabor features which were computed using the log-Mel spectrogram. However, a direct comparison of the two, both with the same compression scheme and power bias subtraction, was not made. A direct comparison of these two for Gabor features is made in [12] on the Aurora-2 corpus. They find, as we will confirm, the performance gain achieved by using a power-law nonlinearity instead of the natural logarithm for dynamic range compression of the spectrogram before spectro-temporal feature extraction. The major differences in these studies and the present work are: we use filter-bank for extracting spectro-temporal features from time-frequency representations and perform dimensionality reduction of the resulting features using the discrete cosine transform (DCT) computed per filter-bank sub-band, rather than with multilayer perceptrons (MLPs) computed on the entire feature set versus . An advantage of the DCT method is that it requires no training.

In our previous work, we introduced Directional Derivative (DD) features: a multi-resolution, spectro-temporal speech representation that filters the log-Mel spectrogram with two dimensional oriented derivative filters [13]. We drew comparisons to both Mel-frequency cepstral coefficients (MFCCs) and Gabor features and demonstrated their competitive accuracy to these existing representations. These features successfully model speech salient information such as energy onset/offset regions and the rising and falling of formants, and do so robustly in the presence of noise. We chose to use the log-Mel spectrogram as the base from which the original DD features were extracted, because it is the most common time-frequency representation used for speech modeling and to give a more direct comparison to other spectro-temporal features. We now explore the efficacy of the Directional Derivative methodology when paired with alternate time-frequency representations.

2. Methodology

We begin by extracting time-frequency representations from speech for comparison. Each of the time-frequency representations is extracted for various number of frequency bins and dynamic range compression schemes. All the evaluated representations are computed using 25 ms frames at a 10 ms rate weighted with a hamming window. Directional Derivative features will then be extracted from each of the resulting representations and tested for word and phone recognition.

2.1. Time-Frequency Representations

2.1.1. Short-time Fourier Transform

The short-time Fourier transform is the most general of the compared time-frequency representations. It makes no assumptions about the human auditory system in addition to frequency analysis. This serves as a comparison of time-frequency representations with non-perceptually inspired versus perceptually inspired processing. It also serves as the base from which the perceptually motivated time-frequency representations will be extracted via non-linear frequency scaling and compression.

2.1.2. Mel-frequency

The Mel-frequency spectrogram is one of the first time-frequency representations to be inspired by human auditory perception. It scales the frequency axis to the Mel scale using overlapping triangular filters. The Mel scale is an approximation to the non-linear scaling of frequencies in the cochlea. This is typically followed by dynamic range compression using the log operator. The Mel-frequency spectrogram is one of the most widely used and it is the basis for Mel-frequency cepstral coefficients, which are a standard feature for many speech recognition systems.

2.1.3. Gammatone

The Gammatone (GT) spectrogram addresses limitations of the Mel-frequency representation. The most significant of these is the introduction of asymmetric filters to replace the triangular filters of the Mel filter-bank [14]. It was argued that these filters better approximate the filtering done in the basilar membrane. The final stage of processing is cubed root compression. The cubed root is motivated by Steven's power-law of hearing [4]. Other power-law non-linearities have been proposed for use with Gammatone based features [15,16], however to our knowledge, a direct empirical comparison has not been made.

2.2. Directional Derivative Features

2.2.1. The Steerable Pyramid Filter-bank

Directional Derivative features are computed by extracting spectro-temporal modulations, using the steerable pyramid wavelet from time-frequency images. The steerable pyramid filter-bank consists of multi-resolution, two-dimentional oriented derivative filters [17]. These filters are designed according to the equation:

$$\psi_k(\vec{\omega}) = (-j\vec{\omega}\cos(\theta - \theta_k))^K \psi(\vec{\omega}), \tag{1}$$

where $\theta = \arg(\vec{\omega})$, $\theta_k = \frac{k\pi}{K}$, k indicates the k^{th} orientation, and K is the derivative order. In this equation, $\vec{\omega} = \left[\begin{smallmatrix}\omega_t \\ \omega_f\end{smallmatrix}\right]$, where ω_t is the frequency of time frames axis and ω_f is the freuency of frequency bins axis, and $\psi_k(\vec{\omega})$ is the mother wavelet prototype. We show the impulse response of the filters used in for extracting DD features for the present work in Figure 1. We use the first and second levels of decomposition and the $45°$, $62.5°$, $90°$, $-62.5°$, and $-45°$ filter orientations, respectively, where the $90°$ filter computes the derivative along the time axis in a similar manner to Δ features. We employ the Steerable Pyramid Toolbox implementation for all experiments [18].

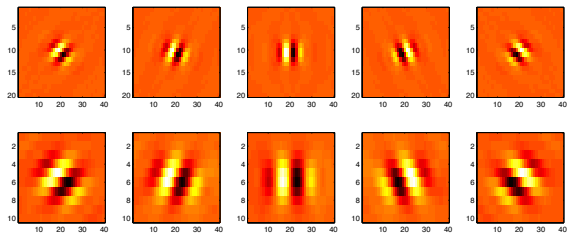

Figure 1: *Impulse response of the steerable pyramid filter-bank.*

2.2.2. Feature decorrelation and dimensionality reduction

We wish to obtain decorrelated feature dimensions, so as to comply with diagonal covariance Gaussian assumptions, which are commonly made to reduce the complexity of the acoustic models. To achieve this goal and for dimensionality reduction we apply the DCT to the output sub-bands of the steerable pyramid filter-bank. Typically, approximately half the coefficients are retained when applying the DCT to a spectrogram to produce cepstral features. This is due to assumptions about the smoothness of the frequency axis of the speech representation. When we extract directional components from speech using oriented filters this assumption no longer holds. We use the relation:

$$M_\theta = \left\lceil \left(1 - \frac{|\theta|}{180°}\right) N \right\rceil, \tag{2}$$

to choose the number of coefficients retained, M_θ, from a particular filter, where θ is the orientation angle, N is the dimension of the filter output, and $\lceil \cdot \rceil$ is the ceiling operator . Further information on Directional Derivative feature computation can be found in [13].

2.2.3. DD features extracted from alternate time-frequency representations

We show the compared time-frequency representations of the phomeme 'iy' from the word 'greasy' in Figure 2. Below each spectrogram is the corresponding sub-band from the $45°$ filter, the sub-band that is responsible for capturing rising formants. This figure provides a depiction of the ability of the steerable pyramid filter-bank to robustly capture formant dynamics in both clean and noise corrupted speech.

3. Experiments and Results

Next, we evaluate the compared features with two automatic speech recognition tasks. The first is continuous digit recognition using the Aurora-2 corpus [19]. The second is continuous phone recognition using the TIMIT corpus [20]. For both tasks, we examine the noise robustness of the compared features with speech corrupted by additive noise for several speech-to-noise ratios (SNRs). We use MFCCs and GFCCs with their first and second temporal derivatives ($\Delta\&\Delta\Delta$) as the first baseline, as these is two of the most common features for ASR. The second baseline is DD features extracted from the log-Mel spectrogram, as this was the original spectrogram used for this task. For all DD features in this work, we concatenate 13 cepstral coefficients computed from the associated time-frequency representation. This yields linear-frequency cepstral coefficients (LFCCs), MFCCs, and Gammatone-frequency cepstral coefficients (GFCCs) for the STFT, Mel-frequency, and Gammatone representations, respectively. This is to provide a static representation to compliment the dynamic DD features.

3.1. Spoken Digit Recognition

The Aurora-2 corpus consists of a training set with 8,440 utterances and three testing sets each with 1,001 utterances. Each utterance is a sequence of the digits 0-9 [19]. The the first two test sets contain four distinct noise types and the third contains two noise types from the other test sets that were collected with a different recording procedure. We use the Hidden Markov Model Toolkit (HTK) [21] to perform recognition. Each word is modeled with a left-to-right Hidden Markov Model (HMM) with 16 states and twenty diagonal covariance Gaussians per state. All models are trained on clean speech and tested on

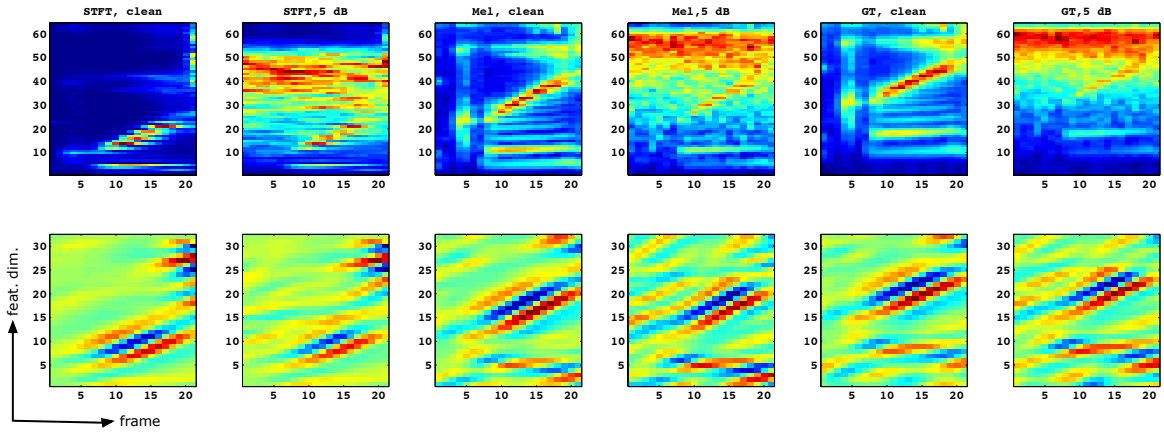

Figure 2: *The top row shows the compared time-frequency representations for clean and 5 dB SNR conditions. The bottom row shows the resulting outputs of the* 45° *steerable pyramid filter-bank sub-band corresponding to the time-frequency representations above.*

speech corrupted with additive noise. We use mean-variance normalization, per utterance, for all final features as it offers additional noise robustness [22].

We show word recognition accuracy (WRA) for the MFCC baseline in Table 1. In Table 2, we compare WRA for DD features extracted from each TFR with a varied number of frequency channels in order to examine how spectrogram frequency resolution affects recognition performance of the resulting spectro-temporal resolution. In this comparison, we use the dynamic range compression scheme which was proposed for each time-frequency representation (i.e., no compression for the STFT, log for Mel-frequency, and cubed root for Gammatone). First, we find that DD features extracted from the log-Mel spectrogram significantly ($p < 0.05$ using the Wilcoxon rank sum test) outperform the baseline MFCC features for all SNRs. We find that 64 frequency bins gives better WRA than 24 or 32 bins for nearly all SNR levels. Both DD of log-Mel and GT outperform the MFCC baseline for all SNRs. DD of GT outperforms GFCCs for 0 and 5 dB SNRs ($p < 0.05$). DD of GT also outperforms that of log-Mel for SNRs 0-15 dB, and the result is significant ($p < 0.05$) for SNRs 0-10 dB, when comparing representations with the same number of frequency bins.

Table 1: *Word Recognition Accuracy (%) of baseline features.*

Feature	20dB	15dB	10dB	5dB	0dB
MFCC	98.4	96.4	91.0	75.6	48.0
GFCC	98.9	97.7	94.1	83.3	58.3

Table 2: *Word Recognition Accuracy (%) of DD features extracted from time-frequency images with various number of frequency bins. We show the highest WRA for each SNR in* ***bold****.*

TFR	NB	20dB	15dB	10dB	5dB	0dB
STFT	24	91.6	90.3	86.8	77.7	60.0
	32	92.3	91.1	87.8	79.0	61.3
	64	92.7	91.6	88.5	80.3	62.5
log-Mel	24	98.7	97.1	93.0	82.5	61.2
	32	98.8	97.4	93.5	83.3	61.5
	64	98.8	97.5	94.2	84.1	61.2
$GT^{0.33}$	24	98.7	97.5	94.1	85.0	65.0
	32	**99.0**	97.8	94.8	86.3	66.7
	64	98.8	**97.9**	**95.3**	**87.5**	**68.9**

We evaluate recognition accuracy using the same compression scheme for all three time-frequency representations in Table 3. We choose cubed root compression for this task as it is used for the Gammatone representation, which gave the highest overall accuracy in the first comparison. Also, all three representations in Table 3 use 64 frequency bins as it gave the highest average performance in the previous experiment. We find that cubed root compression greatly increases the WRA for DD features extracted from the STFT and Mel spectrograms. The cubed root Mel representation significantly ($p < 0.05$) outperforms the log-Mel representation for 0-10 dB. Furthermore, when using cubed root compression, DD features extracted from the Mel spectrogram now surpass performance of those from Gammatone for low SNRs (0-10 dB), although this difference is not significant at the 5% level.

Table 3: *Word Recognition Accuracy (%) of DD features extracted from time-frequency images with cubed root compression. We show the highest WRA for each SNR in* ***bold****.*

TFR	20dB	15dB	10dB	5dB	0dB
$STFT^{0.33}$	**98.9**	97.7	94.5	85.4	65.7
$Mel^{0.33}$	98.8	**97.9**	**95.4**	**87.9**	**69.8**
$GT^{0.33}$	98.8	**97.9**	95.3	87.5	68.9

3.2. Phone Recognition

We use the TIMIT database to evaluate performance for the phone recognition task. The TIMIT corpus consists of 630 speakers, each reading ten sentences [20]. The sentences were carefully constructed to achieve phonetic balance. In order to assess the noise robustness of the features, we corrupt the utterances from the TIMIT database with noises from the Noisex-92 corpus. The Noisex-92 consists of fifteen noise types [23]. We use three for this analysis: pink noise, speech babble, and F16 cockpit noise. Phone recognition accuracy (PRA) is averaged across the three noise types for each SNR level. We use the Kaldi speech recognition toolkit for recognition [24]. We trained three-state monophone GMM-HMMs on clean speech and subsequently tested on noise corrupted speech. There are 16,000 Gaussians distributed between all phone models. Mean-variance normalization is applied to all feature representations per speaker.

Table 4 shows phone recognition performance for the MFCC baseline and Table 5 shows performance for DD features extracted from the compared spectrograms. For these experiments the DD features outperform MFCCs for all SNRs and all time-frequency representations. Similar to the digit recognition experiments, spectrograms with cubed root compression result in better performance than those with log compression. Directional Derivatives of Mel and Gammatone with cubed root compression outperform MFCCs, GFCCs, and their log version for all SNR levels. For both cubed root and log, DD features from Mel spectrograms slightly outperform those from Gammatone for all SNRs, although these differences are not statistically significant.

Table 4: *Phone Recognition Accuracy (%) of baseline features.*

Feature	20dB	15dB	10dB	5dB	0dB
MFCC	64.66	56.87	46.91	36.94	29.25
GFCC	67.55	63.58	56.91	47.30	36.74

Table 5: *Phone Recognition Accuracy (%) of DD features extracted from time-frequency images. We show the highest PRA for each SNR in* ***bold****.*

TFR	20dB	15dB	10dB	5dB	0dB
log-Mel	66.23	60.32	52.01	42.57	33.70
log-GT	65.78	59.33	50.25	40.42	32.43
$\text{STFT}^{0.33}$	66.91	63.02	57.00	48.28	37.49
$\text{Mel}^{0.33}$	**69.16**	**65.82**	**60.39**	**51.94**	**41.32**
$\text{GT}^{0.33}$	68.47	65.05	59.54	51.12	40.35

4. Conclusions and Future Work

We presented an empirical comparison of ASR performance of Directional Derivative features extracted from various time-frequency representations. We demonstrated the word and phone recognition performance differences that result from changing the spectrogram from which spectro-temporal features are extracted. We found that the dynamic range compression scheme had a large effect on performance, with the cubed root giving significantly better performance than the log, which is likely because it more accurately reflects processing of the human auditory system. There was not, however, a larger performance difference between the DD features extracted from perceptually motivated time-frequency representations, i.e., the Mel and Gammatone spectrograms.

In the future, we plan to explore dimensionality reduction of these features with global decorrelation schemes, such as principle component analysis, and discriminative methods, such as linear discriminant analysis. Also, we plan to investigate the efficacy of Directional Derivative features for large vocabulary continuous speech recognition (LVCSR) tasks.

5. Acknowledgements

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

6. References

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

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

[3] 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.

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, p. 1738, 1990.

[5] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

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

[7] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

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

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

[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] 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.

[12] S.-Y. Chang, B. T. Meyer, and N. Morgan, "Spectro-temporal features for noise-robust speech recognition using power-law nonlinearity and power-bias subtraction," in *Proc. ICASSP*, 2013, pp. 7063–7067.

[13] J. Gibson, M. Van Segbroeck, A. Ortega, P. Georgiou, and S. Narayanan, "Spectro-temporal directional derivative features for automatic speech recognition," in *Proc. Interspeech*, 2013.

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

[15] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction." in *Proc. Interspeech*, 2009, pp. 28–31.

[16] ——, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4101–4104.

[17] 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.

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

[19] 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.

[20] J. S. Garofolo, *TIMIT: acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, 1993.

[21] 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.

[22] 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.

[23] A. Varga and H. J. 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.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.