



# Using Spectro-Temporal Features to Improve AFE Feature Extraction for ASR

Suman V. Ravuri<sup>1,2</sup>, Nelson Morgan<sup>1,2</sup>

<sup>1</sup>International Computer Science Institute, Berkeley, CA, USA

<sup>2</sup>EECS Department, University of California - Berkeley, Berkeley, CA, USA

ravuri@icsi.berkeley.edu, morgan@icsi.berkeley.edu

## Abstract

Previous work has shown that spectro-temporal features reduce WER for automatic speech recognition under noisy conditions. The spectro-temporal framework, however, is not the only way to process features in order to reduce errors due to noise in the signal. The two-stage mel-warped Wiener filtering method used in the “Advanced Front End” (AFE), now a standard front end for robust recognition, is another way. Since the spectro-temporal approach can be applied to a noise-reduced spectrum, we wanted to explore whether spectro-temporal features could improve the performance of AFE for ASR. We show that computing spectro-temporal features after AFE processing results in a 45% relative improvement compared to AFE in clean conditions and a 6% to 30% improvement in noisy conditions on the Aurora2 clean training setup.

**Index Terms:** automatic speech recognition, spectro-temporal features

## 1. Introduction

Cortically-inspired spectro-temporal features, which capture spectral and temporal modulations, have successfully been applied to a number of speech recognition and discrimination tasks [12, 14, 5, 9, 18, 19]. In particular, [19] demonstrates that spectro-temporal features perform quite well in automatic speech recognition under noisy conditions. We surmise that the spectro-temporal feature calculation, which filters the log mel-spectra to emphasize many different spectral and temporal modulations, is able to emphasize components of the time-frequency plane that are usable for speech recognition, even if other sections are corrupted. This framework tends to generate many more features than are typically used in ASR, many of which may be highly correlated with one another.

Using redundant features, however, is certainly not the only method to combat noise. An alternative approach is the two-stage mel-warped Wiener filter first introduced in [1] to denoise the input signal prior to feature computation, an improved version of which was later implemented in the “Advanced Front End” (AFE) feature, described in [6]. In addition to a more refined two-stage mel-warped Wiener filter (known simply as “noise reduction” in the original document), AFE includes a waveform processing component that incorporates waveform smoothing and peak picking, a mel-cepstrum calculation, and a blind equalization step. The top pane of Figure 1 represents a vastly simplified diagram of the AFE processing. The bottom pane illustrates the stages of the cepstrum calculation for later reference of the description of the proposed system.

Since AFE processing is complementary to the spectro-temporal one, one could combine much of the AFE processing with spectro-temporal feature extraction. Indeed, this is the method we use in this paper. Figure 2 illustrates this proposed

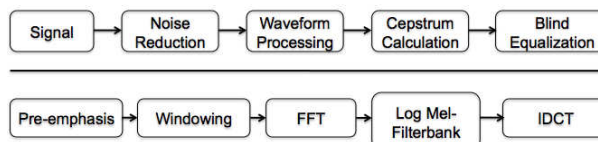


Figure 1: Top Pane: Diagram of AFE feature calculation steps. Bottom Pane: Diagram of the steps in the cepstrum calculation.

feature calculation. We execute the entirety of the noise reduction and the waveform processing steps of the AFE algorithm; then carry out all but the inverse DCT of the cepstrum calculation, thereby leaving us with log mel-spectrogram for use with Gabor filtering; finally, we apply Gabor filtering on the log mel-spectrogram. By incorporating this filtering after execution of AFE processing, we combine two vastly different methods for reducing noise.

## 2. Related Work

Details of this paper fall into two categories: the AFE and spectro-temporal processing. Work on the AFE processing can be traced backed to [1], which first introduced a two-stage mel-warped Wiener filter for reducing noise in the signal. On the Aurora1 data set with clean training and noisy test sets (which contained 4 different noises at 7 noise conditions each), the MFCCs extracted from this denoised signal reduced the average WER by 32.5% relative to MFCCs calculated on the original signal. The AFE in [6], mentioned in the previous section, employed an improved version of the two-stage mel-warped Wiener filter in its processing. AFE’s implementation differs from that of previous work by applying a more sophisticated noise estimation algorithm in the first stage and adding in the second stage an extra step to more aggressively filter purely noisy frames and less aggressively filter speech frames. On the Aurora2 corpus with clean training and noisy test sets, AFE reduced the average WER by 49.1% compared to MFCCs.

The second category of relevant work encompasses spectro-temporal feature processing, typically implemented with Gabor filters. This nascent field has grown in the speech community as Gabor features have shown promise in speech/non-speech discrimination and automatic speech recognition tasks. One challenge in incorporating these spectro-temporal features is that their extremely high dimensionality may cause difficulties in standard HMM-based systems.

Many people have proposed different methods to reduce the dimensionality of spectro-temporal features for different tasks. [12] suggested applying a feature-finding neural network for feature selection. In this algorithm, a neural network

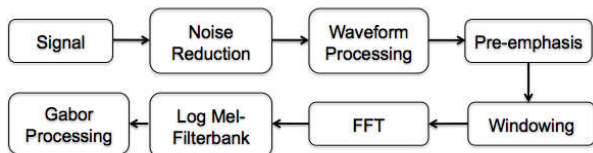


Figure 2: “AFE” and Gabor processing of the input signal.

learns an optimal feature set by replacing an input feature with a randomly-drawn one until the net finds the feature with the smallest increase in classification error. [5] employed a winner-take-most approach which suppresses the least active spectro-temporal neurons in their speech recognition system. A completely different method, used in [14], has been quite successful in automatic speech/non-speech discrimination; Mesgarani et al. extended classical Principal Components Analysis (PCA) to multidimensional tensors in order to reduce feature dimensionality.

An alternate approach partitions spectro-temporal features into different streams, individually processes each stream, and then merges the processed streams prior to inputting them into the decoder. [9, 18, 19, 17] use this approach successfully with spectro-temporal features; each of their systems decrease WER in ASR in clean and noisy conditions. This approach has also been used in speech recognition systems prior to the advent of Gabor features, such as in multi-band systems (see [2]) and in systems incorporating temporal critical bands and PLP (see [16]).

The advantage of using this multi-stream approach is that streams can be divided according to psycho-acoustic and physiological findings. Moreover, since the streams are considered independent, we can process the streams in parallel. The disadvantage of this approach is that the feature dimensionality of the streams must be reduced in some way and a principled way of merging the processed streams must be determined.

The neural-network-based Tandem approach, originally proposed in [8], proves an effective way of reducing the feature dimensionality of the stream. Merging streams, however, still remains an open question. Previous work has considered two different options: weighting the feature streams such that the weights are static across the frames; and weighting the streams dynamically based on either some metric or learning algorithm. [19, 15] weighted streams through inverse entropy and learned weights using a weight-generating multilayer perceptron, while [17] used a hierarchical neural-network-based weighting system to combine features.

### 3. Experimental Setup

For this paper, we use the Aurora2 data set described in [10], a connected digit corpus which contains 8,440 sentences of clean training data and 70,070 sentences of clean and noisy test data. The test set comprises 10 different noises (subway, babble, car, exhibition, restaurant, street, airport, train-station, Modified Intermediate Reference System (MIRS) -filtered subway, and MIRS-filtered street) at 7 different noise levels (clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB), totaling 70 different test scenarios, each containing 1,001 sentences. Since we are interested in the performance of the spectro-temporal features in mismatched conditions, all systems were trained only on the clean training set but tested on the entire test set. The baseline system for this

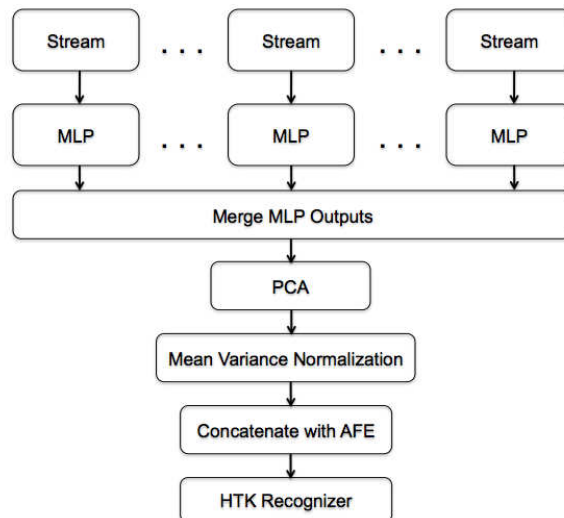


Figure 3: Diagram of processing of the MLP streams.

Feature Stream No.	No. of Features	Spectral Mod.(cyc/chan)	Temporal Mod.(Hz)
1	506	0.04, ..., 0.5 0.04	$\pm 2$ $\pm 4$
2	506	0.13, ..., 0.5 0.04, 0.13	$\pm 4$ $\pm 7$
3	506	0.24, 0.36, 0.5 0.04, 0.13, 0.24	$\pm 7$ $\pm 11$
4	529	0.36, 0.5 0.04, ..., 0.5	$\pm 11$ $\pm 16$

Table 1: Range of spectro-temporal modulation frequencies captured by each of the 4 feature streams.

paper uses only AFE features.

We try two different stream setups, a 4- and 28-stream, as they performed well in previous experiments. The 4-stream systems include streams with different spectral and temporal modulations. The spectral modulations for these streams range from 0.04 cycles per channel to 0.5 cycles per channel while the temporal modulations range from  $\pm 2$ Hz to  $\pm 16$ Hz. Table 1 summarizes the breakdown of the 4 streams.

The 28-stream setup incorporates the aforementioned 4-streams plus 24 others. 16 of these streams are divided along a single temporal modulation (either 2, 4, 6, ..., or 16Hz) and one of two sets of spectral modulations (either 0.1, 0.16, 0.22, 0.28 cycles per channel or 0.34, 0.4, 0.46, 0.52 cycles per channel). The remaining 8 streams are divided along a single spectral modulation (either 0.04, 0.1, 0.16, 0.22, 0.28, 0.34, 0.4, or 0.46 cycles per channel) and a single set of temporal modulations (2, 4, 6, ..., 16Hz). This stream arrangement is detailed in [19].

From this point, the experiments are rather straightforward. For each signal, we perform the standard AFE and Gabor processing as shown in Figure 2. The output of the spectro-temporal processing is segmented into either 4 or 28 streams and each stream is inputted into a multilayer perceptron. The input layer contains 9 frames of context, so the size of the input layer is 9 times the number of features in the stream. The

neural network also contains 160 hidden units, and 56 output targets (each corresponding to an English phone). During training, the MLP weights learned are regularized to zero, as we have noted internally that regularization improves ASR performance slightly.

The outputs of the MLP stream provide an estimate of the posterior probability distribution for phones. We combine each of these phone probability estimates across streams either by equal weighting (for the 4-stream case) or inverse entropy (for the 28-stream case).<sup>1</sup> We then apply Karhunen-Loève Transform to the log-probabilities of the merged MLPs to reduce the dimensionality to 32 dimensions and orthogonalize those dimensions. We then mean and variance normalize the features by utterance. Finally, we append this spectro-temporal feature vector to the AFE feature. The augmented feature vector then becomes the observation stream for the HTK decoder.

The parameters for the HTK decoder are the same as that for the standard Aurora2 setup described in [10]. The setup uses whole word HMMs with 16 states with a 3-Gaussian mixture with diagonal covariances per state; skips over states are not permitted in this model. This is the setup used in the ETSI standards competition, from which AFE was created. More details on this setup are available in [10].

As a final experiment, we extracted spectro-temporal features for the 4-stream setup from the clean training data with white Gaussian noise added at a 15dB SNR. The idea is that training on easily generated noise, even if mismatched to the test conditions, may improve performance in the noisy setting. These Gabor features are appended to an AFE feature vector generated from the clean signal (pilot experiments indicated that using AFE generated with the noisier signal produced worse results).

## 4. Results

Typical results on the Aurora2 test set using the ETSI setup report accuracies (or mean accuracy) across the 10 noises at 7 noise conditions. We do not report accuracies here for two reasons. The first and rather mundane reason is that reporting 280 numbers will result in a table too large for the length constraints of this paper.<sup>2</sup> Second, and perhaps more importantly, we do not think that reporting accuracies in general (even with a reduced table) is properly illustrative of the performance of the system. Consider, for instance, a table consisting of results for two systems in two noise conditions, one clean and one extremely noisy. If the baseline achieved a 98% accuracy rate on the clean test and 3% accuracy on the noisy one, and the proposed system achieved a 99% and 1.9% accuracy on the clean and noisy conditions, respectively, one would clearly choose the latter system as that system reduced over half the errors on the clean test while performing roughly similarly on the noisy one (that is, neither really worked in noise). If we simply look at mean accuracy, however, we see that the baseline actually outperforms the compared system. The reduction in errors corresponds fairly well to the common costs of using a system (for instance, how often a system must retreat to a human operator). For this reason, we report WER results, which for many years have been the standard for most speech recognition tasks.

<sup>1</sup>We also tested inverse entropy weighting for the 4 streams and equal weighting for the 28 streams, but we observed slightly worse results in these instances than in the published results.

<sup>2</sup>The full table, however, is available at <http://www.icsi.berkeley.edu/~ravuri/GaborAFEresults.html>

For this paper, we average WER across noises and report scores for each noisy condition. This type of reporting appears in Table 2, in which we compare the Gabor systems to the AFE baseline. Table 3, summarizes the relative improvement of the proposed systems over the AFE baseline. We also include the relative degradation of results using MFCCs as features, since MFCCs are generally reported as a baseline for this setup. Finally, all results are significant with a p-value of 0.002 using the differences of proportions significance test.

SNR	AFE	AFE+G4	AFE+G4W	AFE+G28
Clean	1.61%	0.88%	1.46%	0.88%
20 dB	2.87%	2.00%	2.09%	2.07%
15 dB	4.56%	3.93%	3.63%	3.65%
10 dB	8.85%	8.10%	7.77%	7.70%
5 dB	19.94%	19.02%	17.90%	18.21%
0 dB	44.14%	41.37%	40.29%	40.33%
-5 dB	74.35%	70.99%	71.39%	69.79%

Table 2: *WER on Aurora2 test set, averaged across noise conditions. G4 corresponds to the 4 Spectro-Temporal streams, trained on clean speech, equally weighted. G4W corresponds to the 4 spectro-temporal streams, trained on white Gaussian noise added clean speech, equally weighted. G28 corresponds to the 28 spectro-temporal streams, trained on clean speech, combined by inverse entropy.*

SNR	MFCC	AFE+G4	AFE+G4W	AFE+G28
Clean	-7.17%	45.34%	9.25%	45.55%
20 dB	-165%	30.39%	27.29%	27.87%
15 dB	-285%	13.79%	20.39%	19.98%
10 dB	-330%	8.49%	12.18%	13.03%
5 dB	-239%	4.59%	10.23%	8.68%
0 dB	-97.7%	6.28%	8.71%	8.64%
-5 dB	-25.82%	4.52%	3.98%	6.14%

Table 3: *Percentage improvement relative to AFE baseline.*

## 5. Discussion

The results for the 4-stream spectro-temporal features are promising. For the 4-stream features trained on clean signal, adding the extra spectro-temporal information to the AFE features removed 45% of the errors in the clean test, 30% on 20dB SNR set, and anywhere from 4.5% to roughly 14% on noisier conditions. This suggests that the proposed features outperform the AFE baseline on the cleaner conditions but do not increase the WER of AFE for noisier speech. Indeed, all settings tested here exhibit at least a slight improvement over the AFE baseline, and in the cleaner conditions, this difference is quite substantial.

For the spectro-temporal features trained on white Gaussian noise, the performance of the system on the Aurora2 test surpasses that of the Gabor streams trained on completely clean speech for moderate noises (5dB to 15dB SNR) but lags behind for cleaner conditions. For AFE features trained on the clean speech with WGN added, results on the test set were 5% to 20% worse relative to the standard AFE baseline.

The 28-stream spectro-temporal features outperform its 4-stream counterpart in all settings except for the 20dB SNR case.

Moreover, the 28-stream spectro-temporal features perform almost as well in noisier conditions as the 4-stream spectro-temporal features trained on white Gaussian noise.

In some sense, this is an unfair comparison for the Gabor features. We are comparing these spectro-temporal features to AFE, which was created for the Aurora2 test set (indeed during the standards competition, the design teams were able to tune on the test set). Furthermore, we are using the same HTK decoder setup that was used for the ETSI standards competition, so we were not able to tune the decoder to achieve the best results. That we are able to achieve any improvement at all on this rather restrictive setup indicates the robustness of this approach.

## 6. Future Work

Despite the promising results of the spectro-temporal features with AFE processing in noise, some open questions still exist. First, limitations of the test set call into question whether these improvements will transfer over to other test sets. Such limitations include: that the noise is added to the test set and that noise is stationary; reverberation is non-existent; and the speech is read connected digits. It would be interesting to see how these features perform on a more realistic ASR task; in particular, the varying temporal properties of the Gabor streams might lend themselves to handling significant variations in speaking rate, as well as other phenomena of conversational speech.

Another challenge is how best to combine MLP outputs. In this paper alone, there are two sub-questions that are interesting for future work. First, since the 28-stream spectro-temporal features include the 4-stream spectro-temporal features, why does the 28-stream spectro-temporal features perform worse than its 4-stream counterpart on the 20dB SNR condition? An optimal combination method would at least be able to choose the 4-streams when the other 24 streams are non-informative. This suggests that more dynamic weighting methods need to be investigated.

We would also like to find a way to combine the 4-stream feature trained on clean speech to that trained on white Gaussian noise to achieve better results. This is a slightly different issue than the one above, because we know that both features contain the same type of information, so it would be interesting to see if there is some method for reducing errors by using clean and noisy features of the same speech.

Finally, in this and previous works, we have appended our Gabor features to either an MFCC or an AFE (which is a processed version of an MFCC). A subset of the Gabor streams, however, contain all the information of an MFCC or in this work, an AFE feature. Without these MFCCs or AFE features, Gabor streams alone perform a few percent worse absolute, suggesting that the features are a suboptimal representation for the standard HMM decoder. We are investigating methods to reduce this mismatch.

Despite these open questions, it appears that the spectro-temporal features can improve ASR over a good baseline in a variety of additive noise conditions.

## 7. Acknowledgements

We would like to thank Adam Janin and Andreas Stolcke for ideas to help make this research better. We would also like to thank the National Defense Science and Engineering Graduate Fellowship (NDSEG) for helping to fund this research. Finally, we would like to thank Sarah Downs for reviewing this work and offering helpful suggestions.

## 8. References

- [1] Agarwal, A., Cheng, Y.M., "Two-stage Mel-warped Wiener Filter for Robust Speech Recognition". The 1999 International Workshop on Automatic Speech Recognition and Understanding, pp. 67-70, 1999.
- [2] Boulard, H. and Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands", In Proc. of Intl. Conf. on Spoken Language Processing, Philadelphia, PA, pp. 422-425, 1996.
- [3] Chi, T., Gao, Y., Guyton, M.C., Ru, P., and Shamma, S.A., "Spectro-temporal modulation transfer functions and speech intelligibility", J. Acoust. Soc. Am., 106(5):2719-2732, 1999.
- [4] Cole, R., Fauty, M., Noel, M. and Lander, T. "Telephone speech corpus development at CSLU", in Proc. Int. Conf. Spoken Lang. Proc., Yokohama, Japan, pp. 1815-1818, 1994.
- [5] Domont, X., Heckmann, M., Joublin, F., Goerick, C., "Hierarchical spectro-temporal features for robust speech recognition", In Proc. ICASSP, Las Vegas, USA, pp. 4417-4420, 2008.
- [6] ETSI standard doc. "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Feature Extraction Algorithm", ETSI ES 202 050 Ver.1.1.1 (2002-10).
- [7] Gelbart, D., "Noisy numbers data and numbers testbeds", International Computer Science Institute, Berkeley, CA. <http://www.icsi.berkeley.edu/speech/papers/gelbart-ms/>.
- [8] Hermansky, H., Ellis, D., Sharma, S., "Tandem connectionist feature extraction for conventional HMM systems", in Proc. ICASSP, Istanbul, Turkey, pp. 1635-1638, 2000.
- [9] Hermansky, H., Fousek, P., "Multi-resolution rasta filtering for tandem-based asr", In Proceedings of Interspeech, Lisbon, Portugal, pp. 361-364, 2005.
- [10] Hirsch, H.G., and Pearce, D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in ISCA ITRW ASR: Challenges for the Next Millennium, Paris, France, pp. 18-20, 2000.
- [11] Kanedera, N., Arai, T., Hermansky, H., Pavel, M., "On the relative importance of various components of the modulation spectrum for automatic speech recognition", Speech Communication, 28:43-55, 1999.
- [12] Kleinschmidt, M., "Localized spectro-temporal features for automatic speech recognition", in Proceedings of Eurospeech, pp. 2573-2576, 2003.
- [13] Lei, X., Siu, M., Hwang, M.Y., Ostendorf, M., and Lee, T. "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition", in Proc. of Intl. Conf. of Spoken Language Processing, Pittsburgh, PA, pp. 1237-1240, 2006.
- [14] Mesgarani, N., Slaney, M., and Shamma, S., "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", IEEE Trans. Audio, Speech, and Language Proc., 14(3):920-929, 2006.
- [15] Misra, H., Boulard, H., Tyagi, V., "New entropy based combination rules in HMM/ANN multi-stream ASR, in Proc. ICASSP, pp. II-741-4 vol.2, Hong Kong, 2003.
- [16] Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivasdas, S., Shinzaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cetin, O., Boulard, H., and Athineos, M., "Pushing the envelope - aside", IEEE Signal Processing Magazine, 22(5):81-88, 2005.
- [17] Valente, H. and Hermansky, H., "On the combination of auditory and modulation frequency channels for ASR applications", In Proceedings of Interspeech, Brisbane, Australia, pp. 2242-2245, 2008.
- [18] Zhao, S.Y., Morgan, N. "Multi-stream spectro-temporal features for robust speech recognition", In Proceedings of Interspeech, Brisbane, Australia, pp. 898-901, 2008.
- [19] Zhao, S., Ravuri, S., and Morgan, N. "Multi-Stream to Many-Stream: Using Spectro-temporal Features for ASR", In Proceedings of Interspeech, Brighton, UK, pp. 2951-2954, 2009.