EASY DOES IT: ROBUST SPECTRO-TEMPORAL MANY-STREAM ASR WITHOUT FINE TUNING STREAMS

Suman V. Ravuri and Nelson Morgan

International Computer Science Institute, Berkeley, CA 94704, USA
University of California - Berkeley, Berkeley, CA 94704, USA

ABSTRACT

Previous work has shown that spectro-temporal features reduce the word error rate for automatic speech recognition under noisy conditions. These systems, however, required significant hand-tuning in order to determine which spectral and temporal modulations should be included in a particular stream. In this work, streams are split into one spectral and temporal modulation each and their posterior probabilities are combined once each stream is discriminatively trained via multilayer perceptron. We show that this combination structure performs as well or better than more elaborate methods in which multiple spectral and temporal modulations are hand-picked per stream. In addition, these type of features outperform standard noise-robust features such as the "Advanced Front End" features, whereas our hand-picked spectro-temporal features do not.

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 [1, 2, 3, 4, 5]. In particular, [5, 6] demonstrate that spectro-temporal features in a Tandem setup (as described by [7]) perform quite well in automatic speech recognition tasks 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.

The problem with existing approaches, especially [5, 6], is that these systems require significant hand-tuning to perform well on ASR tasks. Specifically, the features have to be segmented into streams (i.e., bundled into particular spectral and temporal modulations) prior to discriminative training to be used as features. While the discriminative training does not pose a problem, the modulations included in a particular stream are generally hand-tuned and the structure of streams that perform well for one task may not be optimal for another.

In this paper, we propose an alternative and simpler method of stream segmentation; each stream contains only one spectro-temporal modulation. Figure 1 shows one way of visualizing this operation. Given an auditory spectogram (which is calculated for many features such as MFCC or PLP features), spectro-temporal processing at one particular spectral and temporal modulation yields a "cortical spectrogram", and one can create many cortical spectrograms by filtering the auditory spectrogram at different spectral and temporal modulations. Each of these cortical spectrograms can be considered a separate stream. The advantage of this approach is that it does not require the hand-tuning step of determining which features are to be included in a single stream. Although a priori these cortical spectrograms seem like a poor choice for streams, since certain cortical spectrograms may not be able to classify certain phones, we will show that this setup outperforms an approach by which the temporal modulations and spectral modulations have been hand-picked for strong performance.

2. FEATURES

In this paper, all spectro-temporal filtering is performed on the log mel-spectra (such as the filterbank features generated by HTK). While other methods for generating spectro-temporal receptor fields (STRFs) exist, our approach follows multiplying a complex sinusoid by a Gaussian envelope. The complex sinusoid (with ω_f and ω_t denoting the spectral and temporal modulation frequencies, respectively) is represented as:

$$C(f, t|\omega_f, \omega_t, f_0, t_0) = \exp(i(\omega_f(f - f_0) + \omega_t(t - t_0))$$

while the Gaussian envelope (with parameters σ_f and σ_t denoting the standard deviation in the spectral and temporal

We would like to thank Bernd Meyer for a number of helpful ideas and discussions. We would also like to thank the National Defense Science and Engineering Graduate Fellowship (NDSEG) for helping to fund this research.

axes) is represented as:

$$G(f,t|f_0,t_0,\sigma_f,\sigma_t) = \frac{1}{2\pi\sigma_t\sigma_f}\exp(\frac{-(f-f_0)^2}{2\sigma_f^2}+\frac{-(t-t_0)^2}{2\sigma_t^2})$$

Here, $\sigma_f = \frac{\pi}{\omega_f}$ and $\sigma_t = \frac{\pi}{|\omega_t|}$. Then, given the log mel spectrogram $S(f,t)$ and Gabor filter $F(f,t|\omega_f,\omega_t,\sigma_f,\sigma_t,f_0,t_0) = C(f,t|\omega_f,\omega_t,f_0,t_0)G(f,t|f_0,t_0,\sigma_f,\sigma_t)$, one can calculate the cortical spectrogram $\hat{S}(f,t)$:

$$\hat{S}(f_0,t_0) = \sum_t \sum_f F(f,t|\omega_f,\omega_t,\sigma_f,\sigma_t,f_0,t_0)S(f,t)$$

In practice, the filter is truncated after 3.0 spectral and temporal periods. Since the output to the filter is complex, one can use either the real or imaginary output of the filter. As both real and imaginary parts exhibit desired spectral and temporal modulation characteristics, we use both components. Finally, log mel-spectrogram values are copied at the edges in order to reduce edge effects that would occur if the log mel-spectrogram were padded with zeros.

Since the dimensionality of the output features is extremely high, one cannot simply use the concatenated features as an input to a multilayer perceptron. In our proposed system, the output features are separated into different streams, each of which is a log mel-filterbank processed by only one spectral and temporal modulation. Table 1 shows the range of spectral and temporal modulations used in the 172 streams (which correspond to each spectro-temporal modulation in the table doubled for real and imaginary components). Each stream serves as an input for a multilayer perceptron. The structure of the MLP is as follows: 567 input units - which corresponds to 21 spectral features with first and second derivatives and 9 frames of context, 500 hidden units, and 56 output units (each of which corresponds to an English phone).

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 by inverse entropy. For each stream i, an entropy of the output posteriors at frame f, denoted as $entropy_{if}$ can be calculated. Then, the weight for stream i at frame f, w_{if}, is calculated as:

$$w_{if} = \frac{1/entropy_{if}}{\sum_{j=1}^{n} 1/entropy_{if}}$$

We then apply the Karhunen-Loève Transform to the log-probabilities of the merged MLPs to reduce the dimensionality to 32 dimensions and orthogonalize those dimensions. Next, the features are mean and variance normalized by utterance and finally appended to the MFCC feature. Figure 2 outlines the steps of this process. We will denote this system as the uni-modulation system.

3. EXPERIMENTAL SETUP

The dataset used for this paper is the Numbers95 Corpus described in [8]. The corpus consists of various numeric portions extracted from telephone dialogues of male and female American-English speakers, with the vocabulary size of 32 words. This training set contains 3590 utterances of clean data, totaling roughly 3 hours, while the two test sets each contains 1227 utterances. The first contains only clean data, while the second contains the same utterances with noise added at five signal-to-noise ratios (20dB, 15dB, 10dB, 5dB, and 0dB). The noises used in the test set are from the RSG-10 collection (described in [9]) and include speech babble, factory floor noise, Volvo car-interior noise, F-16 fighter-jet cockpit noise, Leopard tank-interior noise, and Destroyer battleship operations-room-interior noise. The noises are added in same manner as the Aurora2 corpus.

SRI's DECIPHER recognizer is used for the Numbers95 experiment. The MFCCs are vocal tract length and mean and covariance normalized on a per-speaker basis. The recognizer uses gender-independent, within-word triphone Hidden Markov Models (HMMs); cross-word triphone models are not utilized. Moreover, DECIPHER also includes decision-tree clustering of states and genonic mixture tying described

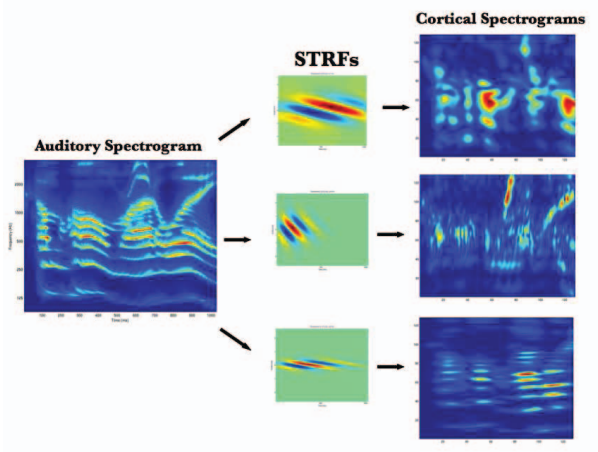

Fig. 1. *Visualization of generating cortical spectrograms from auditory spectrograms.*

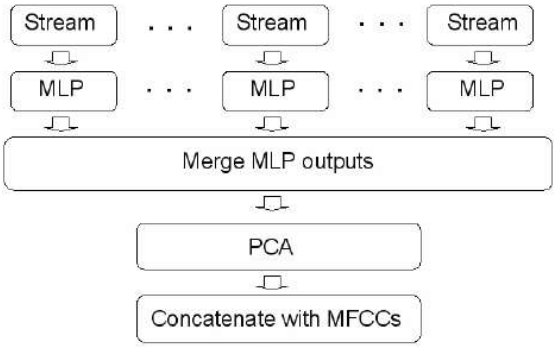

Fig. 2. *Diagram of processing of the MLP streams.*

Spectral Mod. (1/chan)	Temporal Mod.(Hz)
0.04	±6, ±9, ±14.2, ±25, ±50
0.13	±6, ±9, ±14.2, ±25, ±50
0.24	±6, ±9, ±14.2, ±25, ±50
0.36	±6, ±9, ±14.2, ±25, ±50
0.5	±6, ±9, ±14.2, ±25, ±50
0.04, 0.06, . . . , 0.46, 0.48	0
0.00	6, 6.7, 7.7, 9, 8.3, 10, 11.1
0.00	12.5, 14.2, 16.6, 20, 25, 33.3

Table 1. *Range of spectro-temporal modulation frequencies used for the uni-modulation system. Each streams comprises one spectro-temporal modulation (e.g., one stream contains features spectrally modulated at 0.04 chan^{-1} and temporal modulated at 6Hz, another features spectrally modulated at 0.04 chan^{-1} and temporal modulated at -6Hz, etc.) Each modulation generates a real and imaginary component, each of which is separated into different streams.*

in [10].

In this paper, we compare our proposed system to three different baselines. The first uses standard MFCC features with first and second derivatives, while the second is the "Advanced Front End" (AFE) feature proposed in [11], and the third is the 4-stream system proposed in [12] . We use the AFE system as a baseline since it performed well on the Aurora2 noisy digit task, and the 4-stream system since it performed well on the Numbers95 task, shares the same Tandem architecture as the proposed system, and unlike the uni-modulation streams, is hand-tuned for good performance.

The 4-stream system contains the same modulations as the uni-modulation system, but each stream contains features filtered at many different spectro-temporal modulations. Table 2 shows the spectral and temporal modulations included in each stream of the 4-stream system. Each stream in the baseline spectro-temporal system employs a network architecture optimized for the best ASR results. In this work, all streams in the 4-stream system use a 160 unit hidden layer. Although using this network structure results in the number of parameters not being equal to the uni-modulation streams, changing the hidden layer for the baseline systems such that it matches the number of parameters in the uni-modulation system yields significantly poorer performance. Thus, to ensure a fair comparison, we optimize the hidden layer of the baseline systems to give the best recognition results. Finally, the MLP posteriors for the spectro-temporal baseline are combined by inverse entropy and the output posteriors are processed in the same manner as the uni-modulation system described in Section 2.

4. EVALUATION RESULTS

Table 3 displays the results for the three baseline systems - MFCC, AFE, and MFCC + 4-stream spectro-temporal - and the proposed MFCC + uni-modulation spectro-temporal system. The proposed system outperformed the three other baseline systems in both clean and noise-added conditions. This includes AFE, which is specifically designed to reduce errors in a noise-added numbers test setup, and the 4-stream system, whose partitioning of spectro-temporal features were designed specifically to work well on the Numbers95 test set.[1]

Table 5 suggests why the uni-modulation systems outperform streams in which many modulations are used. Despite the fact that individual streams of the uni-modulation system on average perform more poorly on the phone classification

Feature Stream No. *(No. of features)*	Spectral Mod. (1/chan)	Temporal Mod. (Hz)
1 *(462)*	0.04, ... ,0.5	±50
	0.04	±25
	0.04, 0.06, . . . , 0.14	0
	0.00	20, 25, 33.3, 50
2 *(462)*	0.13, ..., 0.5	±25
	0.04, 0.13	±14.2
	0.16, 0.18, . . . , 0.26	0
	0.00	11.1, 12.5, 14.2, 16.6
3 *(462)*	0.24, 0.36, 0.5	±14.2
	0.04, 0.13, 0.24	±9
	0.28, 0.30, . . . , 0.38	0
	0	7.7, 9, 8.3, 10
4 *(462)*	0.36, 0.5	±9
	0.04, ..., 0.5	±6
	0.40, 0.42, . . . , 0.48	0
	0.00	6, 6.7, 7.7

Table 2. *Range of spectro-temporal modulation frequencies captured by each of the 4 feature streams.*

Feature	Clean WER	Noisy WER	Clean Rel. Impr.	Noisy Rel. Impr.
MFCC	2.94%	17.66%	N/A	N/A
AFE	4.65%	15.03%	-58.16%	14.89%
4-stream	2.61%	16.50%	11.22%	6.57%
uni-mod.	**2.52%**	**14.08%**	**14.29%**	**20.27%**

Table 3. *WER on Number95 test set in mismatched conditions. The noisy case is averaged across noise conditions and noise levels from 20dB to 0dB. All specto-temporal systems combine output posteriors using inverse entropy.*

[1]Note that AFE performed better than the 4-stream spectro-temporal system on noise-added test set, but was significantly worse than all systems on the clean condition. To ensure that this result was not a bug in the feature calculation, we ran the same binary on the first four noises of the Aurora test set, and achieved word error rates of .903% and 12.41% on clean and 20-0dB SNR-averaged test sets respectively, which nearly matches the results in [13].

System	Clean Test Phone Accuracy	Noisy Test Phone Accuracy
4-stream	$73.97\% \pm 0.27\%$	$57.52\% \pm 0.08\%$
uni-mod. spectral	$70.31\% \pm 3.87\%$	$52.68\% \pm 4.45\%$

Table 4. *Mean phone accuracy of individual streams for clean and noisy test scenarios of 4-stream and uni-modulation spectro-temporal systems.*

System	Clean Test Phone Accuracy	Noisy Test Phone Accuracy
4-stream	78.13%	60.98%
uni-mod. spectral	80.43%	69.03%

Table 5. *Phone accuracy of combined streams for clean and noisy test scenarios of 4-stream and uni-modulation spectro-temporal systems.*

than those for the 4-stream system (as shown in Table 4), the combined stream for the uni-modulation system is better in both the clean and noisy test conditions than for the 4-stream system. This is especially pronounced in the noise-added test condition, in which the uni-modulation system outperformed the 4-stream system by over 8% absolute. We surmise that while noise may completely corrupt whole streams in the uni-modulation system, there are other streams that are robust to a particular type of noise. By contrast, all streams in the 4-stream system are partially corrupted, leading to generally poorer results in phone and recognition accuracy.

5. FUTURE WORK

Despite the promising results, some open questions remain. One is if these features will work in real noisy situations. The Numbers95 corpus has the limitation that the noise is artificially added, and a number of effects present in real noisy situations, such as the Lombard effect, may change the performance of spectro-temporal features.

Another question is what spectral or temporal features are important for robust ASR. The spectral and temporal modulations included in this paper were picked because such features worked in [5] in a far different stream structure. This does not necessarily imply that these features are optimal for these experiments or adequately cover the range of spectral and temporal modulations needed for robust speech recognition.

Finally, we have combined streams via inverse entropy, but how to optimally combine streams still remains an open problem. It might be best to focus on this problem, as well as on the input features for spectro-temporal processing, rather than worrying about optimizing stream compositions.

6. REFERENCES

[1] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. of Eurospeech*, 2003, pp. 2573–2576.

[2] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *Proc. ICASSP*, vol. 14, pp. 920–930, 2006.

[3] F. Joublin X. Domont, M. Heckmann and C. Goerick, "Hierarchical spectro-temporal features for robust speech recognition," in *Proc. ICASSP*, 2008, pp. 4417–4420.

[4] H. Hermansky and P. Fousek, "Multiresolution rasta filtering for tandem-based automatic speech recognition," in *Proc. of Interspeech 2005*, 2005, pp. 361–364.

[5] S. Ravuri S. Zhao and N. Morgan, "Multi-stream to many-stream: Using spectro-temporal features for automatic speech recognition," in *Proc. of Interspeech*, 2009.

[6] S. Ravuri and N. Morgan, "Using spectro-temporal features to improve afe feature extraction for automatic speech recognition," in *Proc. of ICASSP*, 2010, pp. 1181–1184.

[7] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hidden markov model systems," in *Proc. ICASSP*, 2000.

[8] R. A. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at cslu," in *Proc. of Eurospeech*, 1995.

[9] D. Gelbart, "Noisy numbers data and numbers testbeds," Master's thesis, Univ. of California - Berkeley, Berkeley, CA.

[10] P. Monaco V. Digalakis and H. Murveit, "Genones: Generalized mixture tying in continuous hidden markov model-based speech recognizers," *Proc. ICASSP*, vol. 4, pp. 281–289, 1996.

[11] "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced feature extraction algorithm," 2002.

[12] S. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition," in *Proc. of Interspeech*, 2008, pp. 898–901.

[13] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000, pp. 29–32.