Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR

Sherry Y. Zhao[1], *Suman Ravuri*[1,2], *Nelson Morgan*[1,2]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]EECS Department, University of California at Berkeley, Berkeley, CA, USA
{szhao, ravuri, morgan}@icsi.berkeley.edu

Abstract

We report progress in the use of multi-stream spectro-temporal features for both small and large vocabulary automatic speech recognition tasks. Features are divided into multiple streams for parallel processing and dynamic utilization in this approach. For small vocabulary speech recognition experiments, the incorporation of up to 28 dynamically-weighted spectro-temporal feature streams along with MFCCs yields roughly 21% improvement on the baseline in low noise conditions and 47% improvement in noise-added conditions, a greater improvement on the baseline than in our previous work. A four stream framework yields a 14% improvement over the baseline in the large vocabulary low noise recognition experiment. These results suggest that the division of spectro-temporal features into multiple streams may be an effective way to flexibly utilize an inherently large number of features for automatic speech recognition.

Index Terms: spectro-temporal features, speech recognition

1. Introduction

In order to extract important dynamic information, cortically-inspired spectro-temporal features, which simultaneously capture spectral and temporal modulation frequencies, have recently been used in speech recognition and speech discrimination tasks [1-5]. Despite the number of promising approaches, the utilization and selection of the many possible spectro-temporal features continue to be a challenge, especially for large-vocabulary tasks.

Humans are most sensitive to temporal modulation frequencies up to 16 Hz and spectral modulation frequencies up to 2 cycles per octave [6]. Depending on the desired resolution, it may be possible to have many thousands of filters, each extracting a different combination of spectral and temporal modulation frequencies, while centering on different spectral bands or channels. The relative importance of the features may vary depending on the context. Thus, there is a need to continue exploring the saliency of spectro-temporal-features in different environments as well as methods that allow the dynamic weighting or selection of these features.

2. Related work

There are a number of common methods for the optimal reduction of feature-space dimension, such as is currently required for spectro-temporal features. One such approach is supervised-parameter selection using trained classifiers such as feature-finding neural networks, where optimized features sets are obtained through repeated trials of replacing the feature that leads to the smallest increase in classification error with a randomly-drawn one [1]. Alternatively, a winner-take-most algorithm, where the least-active spectro-temporal neurons are suppressed, has also been successfully employed for automatic speech recognition [2]. A different approach, used for automatic speech-and-non-speech distinction [3], involves multi-dimensional Principal Component Analysis (PCA) through high-order singular value decomposition for decorrelating and reducing the large number of spectro-temporal features.

In addition to these methods, multi-stream approaches, which focus on feature division, have also shown promise in effectively utilizing spectro-temporal features for improving speech recognition performance in both clean and noisy conditions [4, 5, 7]. Multi-stream approaches have been used for some time for speech recognition systems; for instance, in multi-band approaches [8] and for the combination of PLP-based and temporal-based critical band features [9]. In multi-stream approaches, features are generally divided along the spectral and/or temporal axis, consistent with physiological and psycho-acoustical findings from human and other mammalian audition. Advantages to the multi-stream approach are that features may be divided into streams for parallel processing and dynamic weighting or selection, thereby presenting a flexible framework for adapting to the ever-changing acoustic environment.

3. Multi-stream spectro-temporal features for tandem recognition system

3.1. Spectro-temporal feature stream calculation

Spectro-temporal features are extracted from the speech signal using 2-D Gabor filters, employing the method detailed in [1, 5]. The input signal is processed with DC removal, pre-emphasis, Hamming windowing of 25ms in length with 10ms offset, FFT, summation of the resulting squared magnitudes into 23 mel-frequency channels with center frequencies from 124 to 3657 Hz, followed by a log calculation. The number of mel-frequency channels and center frequencies are calculated for telephone speech (8-kHz sampling rate). 2-D Gabor filtering is performed on the resulting log mel-spectrogram. The magnitude of the complex output is taken as the final spectro-temporal feature; in effect, there is one feature per filter, per time frame. Detailed explanations, along with mathematical descriptions of the process and of the Gabor filters can be found in [1].

The 2-D Gabor filters used in this study vary in spectral-modulation frequency, temporal-modulation frequency, and in the mel-frequency channel on which the filter is centered. The spectral-modulation frequencies may range from 0.04 cycles per channel to 0.52 cycles per channel. The temporal-modulation frequencies range from $\pm$2Hz to $\pm$16Hz. These modulation frequency ranges are chosen based on human-

sensitivity data [6] as well as findings on the relative importance of temporal modulation frequencies for automatic speech recognition [10]. The filters may be centered on any of the 23 mel-frequency channels.

In this multi-stream approach, the spectro-temporal features are divided into multiple, parallel streams. Each stream is further processed by a multi-layer perceptron (MLP) that has been discriminantly trained to estimate the posterior probabilities for phone categories, as is commonly done using a Tandem system [11] for recognition experiments. A total of 28 different streams are used in this study. Table 1 contains the range of spectro-temporal modulation frequencies captured by each stream. Three different feature stream divisions methods are employed. Streams 1 through 16 are divided mainly along the temporal modulation domain; each extracts 4 combinations of spectro-temporal modulations, 4 spectral-only modulations, and 1 temporal-only modulation for each mel-frequency channel. Streams 17 through 24 are divided along the spectral modulation domain; each extracts 8 combinations of spectro-temporal modulations, 8 temporal-only modulations, and 1 spectral-only modulation for each channel. The last 4 streams are divided in both the spectral and temporal domains. Streams 25 through 27 contain filters that extract 12 different combinations of spectro-temporal modulations, while Stream 28 extracts 14 different combinations. In addition, Streams 25 through 27 each extracts 4 temporal-only modulations (2-5Hz, 6-9Hz, and 10-13Hz, respectively) and 6 spectral-only modulations (0.04-0.14cyc/chan, 0.16-0.26cyc/chan, and 0.28-0.38cyc/chan, respectively). Stream 28 extracts 3 temporal-only modulations (14-16Hz) and 6 spectral-only modulations (0.4-0.5cyc/chan).

3.2. Feature stream combination

Figure 1 illustrates the usage of MLP-transformed spectro-temporal feature streams in our systems, where the number of outputs for each MLP correspond to 56 phones for the English small vocabulary corpus and 72 phones for the Mandarin-Chinese large vocabulary corpus.

The different sets of phone posteriors may be combined statically or dynamically. One method of static combination is to equally weight each set, effectively taking the average across the streams of phone likelihood outputs. This method was employed in our previous study [5] of a 4-stream framework of spectro-temporal features.

Alternatively, the different sets of MLP outputs may be dynamically combined using stream-specific inverse-Entropy-based weighting [12] or weight generating MLPs. In this study, a single weight-generating MLP is used; inputs consist of 39 Mel-frequency Cepstral Coefficient (MFCC) features, consisting of 13 MFCCs as well as their first and second-order derivatives, and stream-specific frame-level inverse entropies for each frame of the input signal. The weight-generating MLP output for each stream can be viewed as the posterior probability of that stream being the best one. The weight-generating MLP is trained on hard-target labels, where one stream is selected to be the best stream for that frame based on frame-level performance. Ties in frame-level performance are broken by the product of frame-level performance, utterance-level performance, and overall performance of each stream.

Only one type of merging method is used throughout each recognition experiment trial. The outcome should be a single merged phone-probability vector; the vector is decorrelated and its dimensionality is reduced using the PCA transform calculated from the training set.

Feature Stream No.	No. of Features	Spectral Mod.(cyc/chan)	Temporal Mod.(Hz)
1	207	0.1, 0.16, 0.22, 0.28	2
2	207	0.34, 0.4, 0.46, 0.52	2
3	207	0.1, ..., 0.28	4
4	207	0.34, ..., 0.52	4
5	207	0.1, ..., 0.28	6
6	207	0.34, ..., 0.52	6
7	207	0.1, ..., 0.28	8
8	207	0.34, ..., 0.52	8
9	207	0.1, ..., 0.28	10
10	207	0.34, ..., 0.52	10
11	207	0.1, ..., 0.28	12
12	207	0.34, ..., 0.52	12
13	207	0.1, ..., 0.28	14
14	207	0.34, ..., 0.52	14
15	207	0.1, ..., 0.28	16
16	207	0.34, ..., 0.52	16
17	391	0.04	2, ..., 16
18	391	0.1	2, ..., 16
19	391	0.16	2, ..., 16
20	391	0.22	2, ..., 16
21	391	0.28	2, ..., 16
22	391	0.34	2, ..., 16
23	391	0.4	2, ..., 16
24	391	0.46	2, ..., 16
25	506	0.04, ... ,0.5	$\pm$2
		0.04	$\pm$4
26	506	0.13, ..., 0.5	$\pm$4
		0.04, 0.13	$\pm$7
27	506	0.24, 0.36, 0.5	$\pm$7
		0.04, 0.13, 0.24	$\pm$11
28	529	0.36, 0.5	$\pm$11
		0.04, ..., 0.5	$\pm$16

Table 1: Range of spectro-temporal modulation frequencies captured by each of the 28 feature streams.

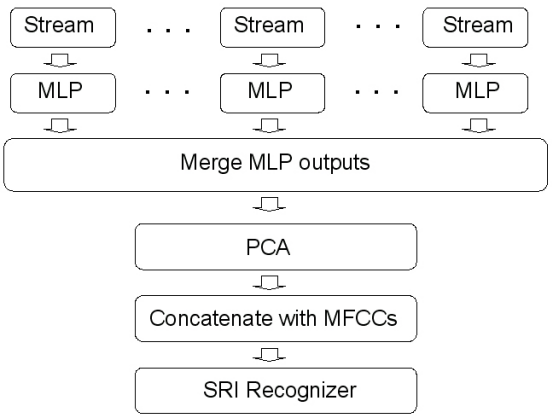

Figure 1: Multi-stream spectro-temporal feature streams for a MLP/SRI-recognizer tandem system.

	Numbers95 Corpus	
	WER	WER
MFCC features (baseline)	2.9%	15.3%
MFCC features + Spectro-temporal features	2.3%	8.1%

Table 2: Comparison of numbers recognition performance of a Tandem system using only MFCC features, consisting of 13 MFCCs and their first and second order derivatives, and MFCC features with multi-stream spectro-temporal features.

	Mandarin Broadcast News/Conversation
	CER
MFCC features (baseline)	25.8%
MFCC features + Spectro-temporal features	22.1%

Table 3: Comparison of Mandarin broadcast news and conversation speech recognition performance of a Tandem system using only MFCC features, consisting of 13 MFCCs and their first and second order derivatives, and MFCC features with multi-stream spectro-temporal features.

4. Small vocabulary task - Numbers95 Corpus

Recognition experiments are conducted on the Numbers95 Corpus [13] using the Tandem recognition system described above. The corpus contains various numeric portions from telephone dialogues of male and female American-English speakers, with a vocabulary of 32 words. The training set for the experiments contains 3590 utterances in clean condition, roughly totaling 3 hours. The testing set contains 1227 utterances, roughly totaling 1 hour; these utterances are exclusive of the training set. There are two experimental conditions for the testing set; one contains all testing-set utterances in clean condition; the other contains the utterances in noise-added conditions. The noise-added test set is created using the principles delineated in the Aurora 2 task [14] using noises of different signal-to-noise ratios from the RSG-10 collection [15].

4.1. Feature streams

A total of 28 streams of spectro-temporal features, as listed in Table 1, are used. The streams are used in the Tandem speech recognition framework, employing a dynamic combination design of merging the stream-specific MLP outputs through a weight-generating MLP, as described in the previous section. The single merged 56-phone-posterior vector is decorrelated using PCA and its dimensionality is reduced to 32 components. A concatenated 71-feature vector, consisting of 39 MFCC features and the 32 spectro-temporal features, serves as input to SRI's DECIPHER. The features are mean and covariance normalized on a speaker basis. The recognizer uses gender-independent, within-word triphone Hidden Markov Models (HMMs); cross-word triphone models are not utilized.

4.2. Recognition task results

Table 2 lists the results of the numbers recognition experiments. Training was conducted on clean Numbers95 utterances for both experiments. For the clean condition, the baseline performance of the Tandem system using 39 MFCC-based features yielded a word-error rate (WER) of 2.9%. Augmenting the features with the spectro-temporal features yielded a word-error rate of 2.3%, a 21% relative improvement on the baseline. A matched-pairs sentence-segment word-error test showed statistical significance with a p-value of less than 0.01, While this is a strong result, we have observed somewhat better performance using a smaller number of streams [5], so the stream weighting method is still not optimum.

For the noisy condition, results are more dramatic. With the same training conducted on clean-condition only utterances, the baseline performance yielded a word-error rate of 15.3% for the noise-added test condition, while the system under test yielded a word-error rate of 8.1%, a 47% improvement on the baseline, a strong result (easily showing statistical significance for p less than 01). This result, using more streams and the MLP-based weighting approach, is notably better than the 30% relative improvement on the baseline for noise-added Numbers95 corpus using 4 equally-weighted spectro-temporal streams, as reported in our previous study [5].

5. Large vocabulary task - Mandarin broadcast news and conversation

For the large vocabulary speech recognition task, experiments are conducted on the Mandarin-Chinese broadcast news corpus used in DARPA GALE evaluations. The vocabulary is roughly 60,000 words. The training data comprises 100 hours sampled from the Mandarin Hub4 (30 hours), TDT4 (89 hours) and GALE Year 1 (747 hours) corpora and includes 50 hours of broadcast news and 50 hours of broadcast conversation data. Finally, the test data for this task is the DARPA GALE 2006 test set[eval06].

5.1. Feature streams

Due to the larger size of this task, a 4-stream system, consisting of features divided along both spectral and temporal modulation domains (Streams 25 through 28 in Table 1) is used for the Mandarin broadcast news corpus. These 4 equally-weighted streams, with quasi-tonotopically divided spectro-temporal features, have been used with promising results in earlier work on the Numbers95 Corpus conducted in this lab [5].

Each stream is trained with an MLP. The MLP input layer for first three streams (Streams 25 through 27 in Table 1) had 4554 units, representing 9 frames of context for a 506-dimensional features, while the input layer for the other stream had 4761 units (since this feature is 529-dimensional). The hidden layer for all four streams are 1150 units. The output layer is 71 phones excluding the reject phone. The log outputs for the four streams are averaged and we use the Karhunen-Loeve Transform (Principle Components Analysis) to reduce the dimensionality to 32.

We combine these features with 13-dimensional MFCCs (vocal tract length normalized) with its first and second derivatives. Moreover, since Mandarin is a tonal language, we included smoothed log-pitch features as described in [16] and first

and second derivatives.

The input features described above are the input to SRI's DECIPHER. The features are mean and variance normalized per speaker. Within-word triphone HMM models are based on a 72-phone model comprising consonants and tonal vowels. Parameters were shared across 2000 states clustered with a phonetic decision tree, and a diagonal-covariance GMM with 32 mixture components modeled the observation distribution. Maximum Likelihood estimation was used to estimate the parameters.

5.2. Recognition results

Table 3 lists the results of the large-vocabulary Mandarin speech recognition experiment. The baseline performance of the Tandem system using 39 MFCCs yielded a character-error rate(CER) of 25.8%. The system under test yielded a CER of 22.1%, a 14% relative improvement on the baseline. The matched-pairs sentence-segment word-error test resulted in a p-value of less than 0.05, indicating a statistically significant improvement. The results are comparable to that obtained using PLP Tandem features, which have been tuned for the task (e.g., for optimal dimensionality reduction).

The relative improvement in performance is lower than that obtained for the Numbers95 corpus. This reduced improvement is similar to what have been observed in other examples of moving techniques from small to large vocabulary tasks. Nevertheless, this preliminary result is quite encouraging, as it can be quite difficult to obtain a greater than 10% reduction in error on tasks such as broadcast Mandarin.

6. Conclusions

Multi-stream spectro-temporal features are utilized for both small and large vocabulary automatic speech recognition tasks. The incorporation of dynamically-weighted spectro-temporal feature streams along with MFCCs yields roughly 21% improvement over the baseline in clean conditions and 47% improvement in noise-added conditions in a small vocabulary speech recognition task; A less elaborate multi-stream framework yields a 14% improvement over the baseline in the large vocabulary task. The results suggest that the multi-stream approach may be an effective way to handle and utilize the potentially large number of spectro-temporal features for speech applications.

7. Acknowledgements

Special thanks to Arlo Faria, David Gelbart, and Adam Janin. Thanks also to Andreas Stolcke and SRI in general for the use of the DECIPHER ASR engine. This research is supported by the IC Postdoctoral Research Fellowship Program. This material is also partly based upon work supported by DARPA under Contract No. HR0011-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of either sponsor.

8. References

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

[2] 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.

[3] 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.

[4] Hermansky, H., Fousek, P., "Multi-resolution rasta filtering for tandem-based asr", In Proceedings of Interspeech, Lisbon, Portugal, pp. 361-364, 2005.

[5] Zhao, S.Y., Morgan, N. "Multi-stream spectro-temporal features for robust speech recognition", In Proceedings of Interspeech, Brisbane, Australia, pp. 898-901, 2008.

[6] 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.

[7] 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.

[8] Bourlard, 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.

[9] Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cetin, O., Bourlard, H., and Athineos, M., "Pushing the envelope - aside", IEEE Signal Processing Magazine, 22(5):81-88, 2005.

[10] 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.

[11] Hermansky, H., Ellis, D., Sharma, S., "Tandem connectionist feature extraction for conventional HMM systems", in Proc. ICASSP, Istanbul, Turkey, pp. 1635-1638, 2000.

[12] Misra, H., Bourlard, 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.

[13] Cole, R., Fanty, 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.

[14] Gelbart, D., "Noisy numbers data and numbers testbeds", International Computer Science Institute, Berkeley, CA. http://www.icsi.berkeley.edu/speech/papers/gelbart-ms/.

[15] 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.

[16] 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.