Multi-Stream Spectro-Temporal Features for Robust Speech Recognition

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

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

Abstract

A multi-stream approach to utilizing the inherently large number of spectro-temporal features for speech recognition is investigated in this study. Instead of reducing the feature-space dimension, this method divides the features into streams so that each represents a patch of information in the spectro-temporal response field. When used in combination with MFCCs for speech recognition under both clean and noisy conditions, multi-stream spectro-temporal features provide roughly a 30% relative improvement in word-error rate over using MFCCs alone. The result suggests that the multi-stream approach may be an effective way to handle and utilize spectro-temporal features for speech applications.

Index Terms: spectro-temporal features, speech recognition

1. Introduction

Speech is rich with acoustic variations that reflect the textures of emotion, context, individuality, and commonality in human vocalization. When separately decomposed into changes in the time and frequency domain, speech may be adequately captured by conventional features used in automatic speech recognition (ASR) systems. For example, spectral fluctuations may be characterized by mel-frequency cepstral coefficients (MFCCs) and perceptual-linear-prediction (PLP) features; temporal fluctuations may be captured by dynamic delta and delta-delta features, as well as by filtering approaches of TRAPS and RASTA [1,2]. Recently, studies in neuroscience have revealed that neurons in the mammalian auditory cortex are highly tuned to specific spectro-temporal modulations [3,4]. These findings have inspired the use of 2-D Gabor filters, which closely resemble the spectro-temporal response fields of neurons [5], to extract features that simultaneously capture spectral and temporal modulation frequencies.

While capturing spectro-temporal modulations may be accomplished through methods such as 2-D Gabor filtering, the selection and utilization of the resulting features pose a somewhat challenging problem. Humans are the 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 Gabor filters, each extracting a different combination of spectral and temporal modulation frequencies, while centering on different spectral bands or channels. If the output of each filter is to be used as a feature for speech recognition or discrimination tasks, then the thousand-plus dimension feature space would most likely be too large to be practical for applications that utilize sizable training data. Furthermore, the relative importance of the features may vary depending on the context. It may be likely that only a subset of the features is extracting salient information at a certain time. Thus, there is a need to explore the saliency of spectro-temporal features in different environments as well as methods that allow the dynamic selection of these features.

2. Related work

Several approaches have been devised to tackle the handling and selection of spectro-temporal features in speech recognition and speech discrimination tasks. A supervised-parameter-selection approach, using feature-finding neural networks (FFNN), was employed by Kleinschmidt to obtain feature sets optimized for different data and target sets in digit recognition tasks [7]. This approach evaluated the importance of a feature by calculating the increase in root-mean-square classification error after the removal of the feature from the set. The least important feature was substituted with a new, randomly-drawn feature. The optimization trials were run over 100 times, resulting in optimized sets of 10 to 80 features. Potential drawbacks to this approach may occur when the data used for the optimization trials insufficiently model the target data. Thus, the selected set of features may not always be optimized if any salient filters are excluded.

A different approach to spectro-temporal feature handling was used by Mesgarani et al. to automatically discriminate speech from non-speech [8]. Instead of selecting a task-optimized set of features, multi-dimensional PCA through high-order singular value decomposition was used to decorrelate and reduce 7680 features to 140 principal components. The minimum number of principal components to achieve 100% classification accuracy was used to determine the number of components to keep. However, performing multi-dimensional PCA directly on spectro-temporal features could lead to limitations such as the loss of non-linear properties and the distortion of the transformation due to outliers that affect the variance.

Furthermore, both studies employ a single-stream approach to selecting and utilizing the spectro-temporal features. Instead of focusing on reducing the number of aggregate features without degrading information representation, this study explores the division of spectro-temporal features into multiple, parallel streams. Each feature stream may be viewed as representing a patch of information in the spectro-temporal response field. This multi-stream approach may be used to systematically analyze feature-stream performance in various environments, as well as enable feature-stream selection.

This paper details the utilization and performance of multi-stream spectro-temporal features in a Tandem system [9,10] using the SRI speech recognizer for numbers recognition tasks. Multi-stream approaches have been used for some time for speech recognition systems, for instance in multi-band approaches [11] and for the combination of PLP-based and temporal-based critical band features [12]. The following sections present the extraction and division of spectro-temporal features, the use of the features in the Tandem system, and the performance of the recognition system using the features in numbers recognition experiments. Furthermore, the relative performance of the feature streams under different noise conditions is discussed.

3. Multi-stream extraction and division of spectro-temporal features

Spectro-temporal features are extracted from speech input using 2-D Gabor filters, employing the method detailed in [7]. As illustrated in Figure 1, the input signal goes through a process consisting of DC removal, pre-emphasis, Hamming windowing of 25ms in length with 10ms offset, FFT, summation of the resulting magnitude into 23 mel-frequency channels with center frequencies from 124 to 3657 Hz. The number of mel-frequency channels and center frequencies are calculated for telephone speech (8-kHz sampling rate) and may be varied depending on the type of speech and sampling rate. Logarithmic compression is performed on the amplitude values. 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 [7].

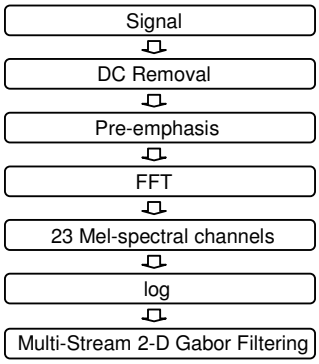

Figure 1. Extraction of spectro-temporal features from speech signal, using 2-D Gabor filters.

The 2-D Gabor filters used in this study vary in spectral-modulation frequency (ω_f), temporal-modulation frequency (ω_t), and in the mel-frequency channel on which the filter is centered. The spectral-modulation frequencies range from 0.04 cycles per channel to 0.5 cycles per channel (approximately 0.14 cycles per octave to 1.6 cycles per octave, with the 23 channels covering 7 octaves); this roughly translates to 1 cycle covering 2 channels to 1 cycle covering all 23 channels. The temporal-modulation frequencies range from 2Hz to 16Hz and -2Hz to -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 [13]. The filters may be centered on any of the 23 mel-frequency channels.

Figure 2 illustrates the division of the 2-D Gabor filters that simultaneously capture spectral and temporal modulation frequencies used in this study.

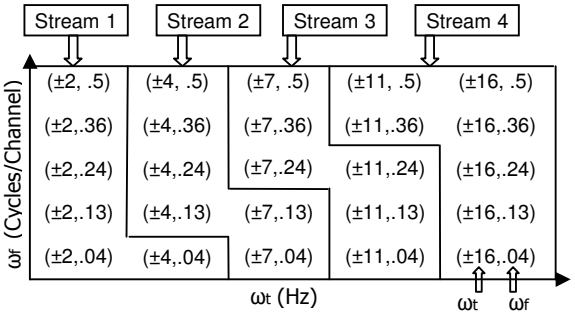

Figure 2: Division of spectro-temporal features into four separate feature streams.

Each stream also contains Gabor filters that capture only the temporal modulation frequency and those that only capture the spectral modulation frequency. Table 1 lists the division of such Gabor filters. It is noted that the spectral modulation frequency is set to 0 for Gabor filters that extract only temporal modulation frequency. The temporal modulation frequency is set to 0 for Gabor filters that extract only the spectral modulation frequency. For each time frame, the Gabor filters are centered on each of the 23 mel-frequency channels. Thus, there are a total of 2047 features. Stream 1, 2, and 3 each has 506 features while Stream 4 has 529 features. The feature-stream division may be viewed as quasi-tonotopic, with Stream 1 consisting of filters capturing the lowest spectral and temporal modulation frequencies and Stream 4 consisting of filters capturing the highest modulation frequencies.

	Temporal Modulation Only ($\omega_f = 0$)	**Spectral Modulation Only** ($\omega_t = 0$)
Stream 1	ω_t={2, 3, 4, 5 Hz}	ω_f={.04, .06, .08, …, .14 cyc/chan}
Stream 2	ω_t={6, 7, 8, 9 Hz}	ω_f={.16, .18, .2, …, .26 cyc/chan}
Stream 3	ω_t={10, 11, 12, 13 Hz}	ω_f={.28, .3, …, .38 cyc/chan}
Stream 4	ω_t={14, 15, 16 Hz}	ω_f={.4, .42, .44, …, .5 cyc/chan}

Table 1: Division of temporal-modulation features and spectral-modulation features into four streams.

4. ASR experiments

4.1. Training and testing data

Recognition experiments were conducted on the Numbers95 Corpus [14] using a Tandem recognition system. The corpus

contains various numeric portions, such as zip codes and street numbers, extracted from thousands of telephone dialogues. In addition, this corpus contains both male and female speakers of different ages and from different dialect regions.

The training set for the experiments contains 3590 utterances in clean condition. The testing set contains 1227 utterances that 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 was created using the principles delineated in the Aurora 2 task [15]. The noises used in the noise-added test set are selected from the RSG-10 collection; they 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 [16]. Five signal-to-noise ratios, ranging from 0dB to 20dB, along with the no-noise-added condition are used.

4.2. Multi-stream spectro-temporal features in a Tandem recognition system

Figure 3 illustrates the usage of spectro-temporal features in a MLP/conventional-speech-recognizer Tandem system [10] to carry out the recognition experiments. The four spectro-temporal-feature streams are each fed into a multi-layer perceptron (MLP) that is trained on the training set, using the same features in the respective stream. The output of each MLP is a vector of 56 phone probabilities. For each time frame, the MLP outputs are merged by taking the log average of the 56 phone probabilities. The merged feature vector is decorrelated using the PCA transform calculated using the training set. The dimensionality of the transformed vector is reduced by keeping 32 principal components. The resulting 32-dimension feature vector is concatenated with 39 Mel-Frequency Cepstral Coefficients (MFCCs) (including first and second derivatives) extracted from the same testing data. The concatenated 71-feature vector serves as input to the SRI speech recognizer that is trained with clean Numbers95 utterances. The baseline for the recognition experiments uses only the 39 MFCCs, without concatenation with other types of features, in the Tandem system.

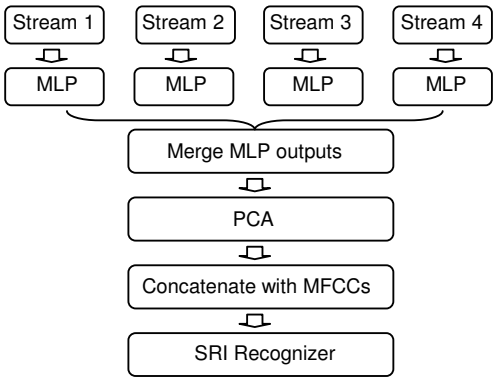

Figure 3. Multi-stream spectro-temporal feature streams for a MLP/SRI-recognizer Tandem system.

4.3. Recognition task results

Table 2 lists the results of the numbers recognition experiments. Training was conducted on clean Numbers95 utterances for both experiments. When the testing set consisted of utterances in the clean condition, the baseline performance of the Tandem system using 39 MFCCs yielded a word-error rate of 2.9%. The performance of the system using MLP outputs obtained using four streams of spectro-temporal features along with the 39 MFCCs yielded a word-error rate of 2.0%, a 31% relative improvement on the baseline. A matched-pairs sentence-segment word-error test resulted in a p-value of 0.001, indicating statistical significance in the performance difference.

	Clean Condition WER	Rel.Imp.	**Noise-added Condition** WER	Rel.Imp.
39 MFCCs (baseline)	2.9%	0.0%	15.3%	0.0%
39 MFCCs + Multi-Stream Spectro-Temporal Features	2.0%	31.0%	10.3%	32.7%

Table 2. Comparison of numbers recognition performance of a Tandem system using MFCCs only and MFCCs with multi-stream spectro-temporal features. Word-error rates (WER) and relative improvements (Rel.Imp.) on the baseline are listed.

When the testing set consisted of utterances in noise-added conditions, the baseline performance of the Tandem system using 39 MFCCs yielded a word-error rate of 15.3%. The performance of the system using MLP outputs obtained using four streams of spectro-temporal features along with the 39 MFCCs yielded a word-error rate of 10.3%, a 32.7% relative improvement on the baseline. A matched-pairs sentence-segment word-error test resulted in a p-value of less than 0.001, indicating statistical significance in the performance difference.

In both experiments, the incorporation of the multi-stream spectro-temporal features in the Tandem system led to roughly a 30% relative improvement on the word-error-rate of the baseline systems. This result suggests that multi-stream spectro-temporal features may be used effectively along with MFCCs in a Tandem recognition system for improved performance.

4.4. Stream-specific MLP performance

An advantage to the Tandem system is that the MLP outputs may be analyzed for insights into the relative performance of each feature stream. This information may be used in developing different feature-stream division methods, as well as devising feature-stream selection or weighting systems. A relative-goodness measure can be derived from the percentage of frames where a stream yields the highest probability among all streams for the labeled phone. Table 3 lists this measure for each of the four streams with respect to the noise condition of the test utterances.

The four streams performed roughly on par with one another under the clean condition, with Stream 3 having a slightly

higher rate of getting chosen as the "best stream". A similar pattern may be seen in the babble-added condition. This finding is not surprising given that both contain speech and speech-like sounds. Among the noisy conditions, Stream 2 consistently had the lowest "best-stream" rate. This observation seems to be the only consistent trend in all of the noise-added conditions, while being absent in the clean condition. The data also reveal that no single stream stands out as a particularly poor performer; the highest and lowest "best-stream" rates do not differ by more than 11% under any condition. These findings suggest that none of the streams should be disregarded under any of the conditions examined.

	Stream 1	Stream 2	Stream 3	Stream 4
Clean	23%	25%	28%	23%
Babble	24%	23%	28%	24%
Factory floor	24%	22%	25%	28%
F16 cockpit	27%	21%	29%	23%
Volvo interior	27%	21%	26%	26%
Leopard tank interior	25%	19%	26%	30%
Destroyer Ops interior	26%	21%	27%	26%

Table 3. Percentage of frames where a stream yields the highest probability for the labeled phone, with respect to the noise condition of the test utterances.

Future extensions of this research may focus on analyzing the performance spectro-temporal features streams under different acoustic conditions. Stream-specific MLP outputs may be used to gauge the relative-goodness of the streams. The results of such analysis may be used to develop and implement feature-stream selection and/or weighting systems. It would also be fruitful to explore different methods of dividing the spectro-temporal features. The number of streams, the number of features per stream, and the grouping methodology of the features are all factors that may be varied in the implementation of multi-stream spectro-temporal features to achieve robust performance in speech applications.

5. Conclusions

The approach of multi-stream division, rather than feature reduction, to handling the large number of spectro-temporal features was investigated. The feature-stream division implemented here is quasi-tonotopic, with Stream 1 consisting of filters that capture the lowest spectro-temporal modulation frequencies and Stream 4 consisting filters that capture the highest spectro-temporal modulation frequencies. The incorporation of multiple streams of spectro-temporal features along with MFCCs in the Tandem system has resulted in improved performance in number recognition tasks in clean and noisy environments. 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. Extensions of this study may involve the analysis of feature-stream performance under various acoustic conditions, as well as implementing feature-stream selection and different methods of feature division. Finally, as the methods mature, they can be applied to large vocabulary tasks such as the ones that have been handled with previous multi-stream methods, e.g., for DARPA tasks including EARS and GALE.

6. Acknowledgements

Special thanks to Arlo Faria, David Gelbart, Adam Janin, Kofi Boakye, and Gerald Friedland. This research is supported by the Intelligence Community Postdoctoral Research Fellowship Program.

7. References

[1] Hermansky, H. and Sharma, S., "Temporal Patterns (TRAPS) in ASR of Noisy Speech", in Proc. ICASSP, Phoenix, Arizona, USA, 1999.

[2] Hermansky, H. and Morgan, N., "RASTA Processing of Speech", IEEE Trans. Speech and Audio Proc., 2(4):578-589, 1994.

[3] Depireux, D.A., Simon, J.Z., Klein, D.J., and Shamma, S.A., "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex', J. Neurophysiology, 85:1220-134, 2001.

[4] Klein, D.J., Depireux, D.A., Simon, J.Z., Shamma, S.A., "Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design", J. Comp. Neuroscience, 9:85-111, 2000.

[5] De-Valois, R. and De-Valois, K., "Spatial Vision", Oxford U.P., New York, 1990.

[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):719-2732, 1999.

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

[8] 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.

[9] Ellis, D., Singh, R., and Sivadas, S., "Tandem acoustic modeling in large-vocabulary recognition", in Proc. ICASSP, Salt Lake City, Utah, USA, 2001.

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

[11] 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, pages 422-425, Philadelphia, October 1996.

[12] 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, Sep. 2005.

[13] 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.

[14] Cole, R., Fanty, M., Noel, M. and Lander, T. "Telephone speech corpus development at CSLU," in Proc. Int. Conf. Spoken Lang. Proc., 1994.

[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, 2000.

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