SPECTRO-TEMPORAL GABOR FEATURES FOR SPEAKER RECOGNITION

Howard Lei, Bernd T. Meyer, and Nikki Mirghafori

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA

ABSTRACT

In this work, we have investigated the performance of 2D Gabor features (known as spectro-temporal features) for speaker recognition. Gabor features have been used mainly for automatic speech recognition (ASR), where they have yielded improvements. We explored different Gabor feature implementations, along with different speaker recognition approaches, on ROSSI [1] and NIST SRE08 databases. Using the noisy ROSSI database, the Gabor features performed as well as the MFCC features standalone, and score-level combination of Gabor and MFCC features resulted in an 8% relative EER improvement over MFCC features standalone. These results demonstrated the value of both spectral and temporal information for feature extraction, and the complementarity of Gabor features to MFCC features.

Index Terms— Speaker recognition, Gabor features, ROSSI database, spectral and temporal modulation

1. INTRODUCTION

The 2D Gabor features have been a more recent development in speech processing applications. They were developed to model certain stimuli to which the neurons of the mammalian auditory cortex are sensitive. These stimuli consist of both spectral and temporal modulation frequencies [2]. Different neurons are sensitive to stimuli of different temporal and modulation frequencies, and many stimuli span more than 200 ms temporally, which far exceeds the span of typical acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs). Because 2D Gabor functions (or filters, that span both the spectral and temporal dimensions) are found to successfully model such stimuli [3], systems based on such Gabor filters attempt to emulate the underlying signal processing strategies of the mammalian auditory system. These Gabor-based systems are able to well-perceive human voices even in the presence of channel and environmental noise.

2D Gabor features were first used by Kleinschmidt et al. in 2002 for automatic speech recognition (ASR), with considerable ASR improvements [4]. Recently, they were used to improve the robustness and word error rate of ASR systems, especially in adverse acoustic conditions [5][6]. While Gabor features have been successfully applied to speaker identification (using a set of 26 speakers) [7], to the best of our knowledge, they have yet to be applied using larger databases in the NIST SRE framework. Furthermore, recent developments in Gabor feature extraction led to further improvements to Gabor feature-based ASR systems [5][6].

In this work, we have attempted to employ 2D Gabor features for large-scale speaker recognition in the NIST SRE framework, and investigated some recent developments in Gabor feature extraction. Our goal was to determine if a fixed set of 2D Gabor filters would respond differently to voices from different individuals, resulting in speaker discriminativeness of the filter responses, and whether or not Gabor features could complement MFCC features. This work mainly focused on handling the number of variables involved in Gabor feature extraction, which could potentially produce Gabor features sets that are significantly different from one another. For example, because of the large sets of spectral and temporal modulation frequencies (typically well over 1,000, and potentially infinite) that could be used for Gabor feature extraction [6], dimensionality reduction was important to ensure that the final set of Gabor feature dimensions were not overly redundant with one another. While such dimensionality reduction techniques have been streamlined for ASR, we investigated such techniques for speaker recognition. We explored the Gabor features on traditional GMM-UBM [8] and GMM-SVM [9] speaker recognition approaches, as well as the i-vector approach. The Gabor features were explored for both the noisy ROSSI database, as well as a subset of the SRE08 male telephone database. Because the ROSSI database lacked sufficient development data for the i-vector approach, only the GMM-UBM and GMM-SVM approaches were applied to the ROSSI database.

This paper is organized as follows: Section 2 describes the databases used, Section 3 describes the 2D Gabor features, Section 4 describes the experiments and results, and Section 5 provides a discussion and summary of our work.

2. DATASET

The ROSSI database, which contains various types of channel and environmental noise typically with 10 dB SNR per conversation side (or utterance), was first used in this work. The ROSSI database conversation sides consist of roughly 50 seconds of monologue landline or cellular phone speech, recorded in various noisy environmental conditions. The breakdown of the development, training and testing conversation sides are shown in table 1.

A total of 200 speakers were used. 100 of those speakers were used to train speaker models, and the other 100 contributed only to impostor trials. Amongst the 200 speakers, 55% were male while 45% were female. There were a total of approximately 2,000 true speaker trials, and 410,000 impostor trials. Amongst the trials, 50% were gender-matched, 67% of the trials were channel-matched, 29% were condition-matched, and 20% were channel- and condition-matched.

Gabor feature performance on the NIST SRE08 male telephone database (with 1,600 conversation sides, 500 speakers, 12,000 trials, and 1,200 true speaker trials) was also investigated. 90 speakers and 900 conversation sides from NIST SRE04 were used as development data. All conversation sides were $\sim$2.5 minutes long, containing speech from one speaker only.

Development			
Environmental Condition	Channel	# of conv. sides	# of hours
Office	Landline	200	2.8
Office	Cellular	50	0.7
Public place	Cellular	50	0.7
Vehicle	Cellular	50	0.7
Roadside	Cellular	50	0.7
Total	–	400	5.6
Training and testing			
Environmental Condition	Channel	# of conv. sides	# of hours
Office	Landline	300	4.2
Office	Cellular	300	4.2
Public place	Cellular	300	4.2
Vehicle	Cellular	300	4.2
Roadside	Cellular	150	2.1
Total	–	1,350	18.9

Table 1. Description of channel and environmental condition information of conversation sides in the ROSSI database.

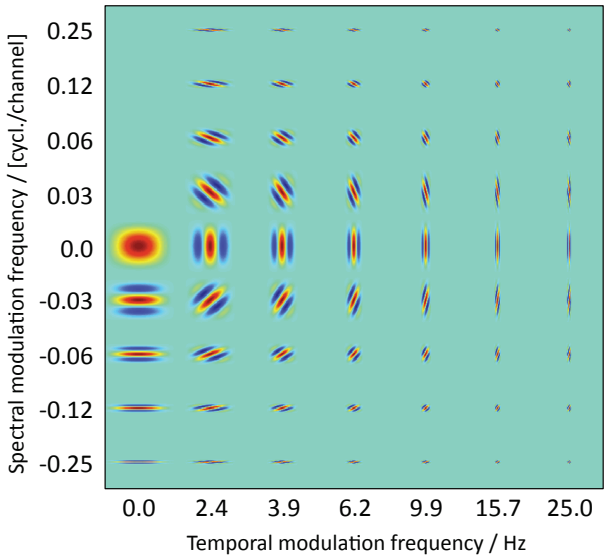

Fig. 1. Real components of the set of 59 2D Gabor filters.

3. GABOR FEATURE EXTRACTION

In this section, we describe the characteristics of the Gabor features. Feature extraction was based on the approach described in [6], one of the most recent state-of-the-art approaches that resulted in the successful application of Gabor features to noise-robust ASR. The features were first calculated by convolving the log mel spectrogram of speech with a set of 2D Gabor filters. Each Gabor filter $g(n,k)$ is a product of a complex sinusoid $s(n,k)$ with a Hann envelope function $h(n,k)$ (the Gabor filters are hence complex functions), defined as follows:

$$s(n,k) = exp[i\omega_n(n-n_0) + i\omega_k(k-k_0)]$$
$$h(n,k) = 0.5 - 0.5cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right) cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right)$$

The ω_n and ω_k terms represent the time and frequency modulation frequencies of the complex sinusoid, while W_n and W_k represent time and frequency window lengths of the Hann window. We used the same set of 59 2D Gabor filters as used in [6], first proposed in [10]. The set of filters were selected to cover a wide range of modulation frequencies, and for their transfer functions to exhibit constant overlap in the modulation frequency domain, which approximated orthogonal filters thereby limiting the redundancy of the filter output. The filter bank parameters (e.g., filter spacing, lowest and highest modulation frequencies in time and frequency dimension) were determined empirically based on a speech recognition task in [10]. In that work, the lowest non-zero temporal modulation frequency was 6 Hz. A slightly modified set of parameters that also covered modulations between 2-4 Hz [6] was used, because frequencies in this range were often found to be important in speech-related tasks. Figure 1 illustrates the 59 Gabor filters.

For each feature frame, each 2D Gabor filter was convolved with a set of 23 log mel spectrum frequency bands, with frequencies ranging from 64 to 400 Hz, producing 59x23=1,357 initial feature dimensions. Because the 2D Gabor filters were complex, the 1,357 feature values were hence complex, and the *real* components of the feature values were used (this resulted in superior ASR performance, as shown in [6]). The 1,357 *real* feature dimensions were reduced to 449 through selective sampling of filter outputs, producing a 449-dimensional feature vector, which was still too large for standard speaker recognition modeling approaches to handle effectively.

The 449-dimensional Gabor features were reduced to a final set of 32 dimensions using a Multi-Layer Perceptron (MLP), followed by a log transformation, followed by Principle Component Analysis (PCA). The inputs to the MLP were the 449-dimensional features, and the outputs were a set of MLP output posteriors of an intermediate dimension. The top 32 PCA eigen-dimensions of the MLP output posteriors were retained, resulting in the final set of 32 dimensions. The reason that PCA was used on top of MLP dimensionality reduction was to obtain a set of orthogonal output feature dimensions (because the MLP output dimensions represented posterior probability distributions, they were correlated with one another). Figure 2 illustrates a simplified view of this process.

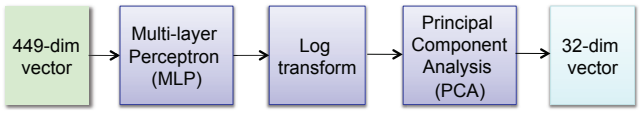

Fig. 2. Simplified view of the Gabor feature dimensionality reduction approach.

Keeping the dimensionality reduction framework, we explored different approaches to MLP training in an attempt to obtain Gabor features tuned for the speaker recognition task. The first approach was to simply use a pre-trained MLP (from the work of [6]) with 56 phone posteriors as outputs, trained on the Aurora2 database. Two additional MLP training approaches were investigated. These approaches used speaker classes as MLP output labels, determined via the clustering of the development data's supervectors. The supervectors consisted of the GMM mean parameters obtained from MAP adaptation of a 128-mixture UBM to each of the conversation sides, using MFCC features C0-C12 + Δ + $\Delta\Delta$. The first clustering approach involved K-means clustering of the supervectors, while the

second involved bottom-up hierarchical clustering. In these two approaches, each development conversation side was associated with one speaker label for MLP training. Both clustering approaches clustered the development conversation sides into a set of 56 speaker clusters, to be consistent with the pre-trained MLP for ASR.

Finally, we investigated using the MLP training labels derived from the UBM mixture likelihoods, where the label for each frame corresponded to the UBM mixture with the highest likelihood (assuming that all mixtures had equal prior). For this approach, the effects of varying the number of UBM mixtures were examined, and the number of MLP output dimensions varied according to the number of UBM mixtures.

4. EXPERIMENTS AND RESULTS

Speaker recognition experiments for the Gabor features were initially run using the well-established GMM-UBM and GMM-SVM speaker recognition approaches. Because of the relatively small size of the ROSSI database, 128-mixture GMM models were used for all system implementations, with mean- and variance-normalized MFCC features C0-C12 + Δ + $\Delta\Delta$. The GMMs were trained using the open-source ALIZE toolkit [11], the SVMs were implemented using the SVM^{light} toolkit [12] (with wrapper scripts from SRI), and MFCC features were extracted using HTK [13]. The SHOUT speech/non-speech detector [14] (which trains a speech, non-speech, and silence model per conversation side) was used to extract the speech regions of each conversation side. The Gabor- and MFCC-based GMM-SVM systems were combined at the feature level. For experiments using the NIST SRE08 dataset (where SRE04 was used as development data), the performance of the Gabor features using the i-vector approach, as described in [15], was also investigated.

Experiments were performed using each of the MLP training approaches described in Section 3. The approach using the pre-trained MLP on the Aurora2 database, with phone posteriors as outputs, is denoted as *phone-mlp*; the approach using k-means clustering for speaker class-based MLP training is denoted as *kmeans-mlp*; the approach using bottom-up hierarchical clustering is denoted as *botup-mlp*, and the approach using UBM likelihoods is denoted as *ubmllk-mlp*. For the latter approach, the optimal speaker recognition performance was obtained using a 76-mixture UBM to extract the Gaussian labels, resulting in 44 distinct labels. Hence, the Gabor feature dimensions were first reduced from 449 to 44 via the MLP, then to 32 via PCA. Table 2 shows the experimental results on the ROSSI database, for all features and systems. Note that the GMM-UBM approach was used to determine the optimal MLP training method.

Feature	MLP training	System	EER (%)
MFCC	–	GMM-UBM	9.2
Gabor	*phone-mlp*	GMM-UBM	17.7
Gabor	*kmeans-mlp*	GMM-UBM	20.7
Gabor	*botup-mlp*	GMM-UBM	21.7
Gabor	*ubmllk-mlp*	GMM-UBM	9.5
Gabor+Δ	*ubmllk-mlp*	GMM-UBM	9.2
MFCC	–	GMM-SVM	7.6
Gabor+Δ	*ubmllk-mlp*	GMM-SVM	7.6
Gabor+Δ + MFCC	*ubmllk-mlp*	GMM-SVM	7.2

Table 2. Gabor feature results on the ROSSI database, for different feature implementations (based on MLP training variants), and speaker recognition approaches. The last row shows feature-level combination.

According to the results, the *ubmllk-mlp* approach labeling method produced the lowest EER among all the MLP-training approaches for the GMM-UBM system: 9.5% EER for Gabor features standalone, and 9.2% EER for Gabor+Δ features. The Gabor+Δ features achieved the same EER (9.2%) as the MFCC baseline. Using the pre-trained MLP based on phone posteriors (*phone-mlp* approach) produced a 17.7% EER, which was lower than the EERs (20.7% and 21.7%) that resulted from the two clustering-based approaches (*kmeans-mlp* and *botup-mlp*, respectively). The results indicated a high EER variability associated with the MLP-based dimensionality reduction approaches, as the methods by which the final 32 dimensions were obtained had large impacts on speaker recognition performance, and would warrant further investigation. The results also suggested that the Gabor features were able to perform well in the presence of channel and environmental noise, present in the ROSSI database.

Feature-level combination of the Gabor+Δ and MFCC features was also investigated using the GMM-SVM system, in which the MFCC supervectors were concatenated with the Gabor+Δ supervectors. Here, the GMM-SVM approach was used over the GMM-UBM approach, because the former could more successfully handle combinations of higher-dimensional features. The Gabor+Δ and MFCC combination produced a 7.2% EER, a 5% relative EER improvement over both the Gabor+Δ and MFCC GMM-SVM systems standalone, as shown in table 2.

Because of the large dimensionality of the Gabor features, it was difficult to test the features on large NIST SRE datasets, due to large amounts of computational time and space required to process the Gabor features. This was especially true for the i-vector with LDA and WCCN approach, because of the large amounts of i-vector development data (at least 10,000 conversation sides per gender) required. We thus obtained some Gabor feature results using an i-vector system, on a subset of NIST SRE08 male telephone data (as described in Section 2), using a set of 90 speakers and 900 conversation sides from SRE04 for i-vector development (including T-matrix, LDA matrix, and WCCN matrix training). We used the same i-vector development data to implement an MFCC-based i-vector system for comparison with the Gabor-based system, but because of the lack of i-vector development data, the MFCC-based results suffered in terms of EER. Gabor feature extraction was performed using the MLPs trained on the ROSSI data, because the ROSSI data provided more optimal MLP training, in terms EER, than SRE data. This was because when the *ubmllk-mlp* approach was applied for MLP training on SRE data, only a few of the Gaussian mixtures had the highest likelihoods for all frames. Table 3 shows the SRE08 results.

Feature	MLP training	System	EER (%)
Gabor+Δ	*ubmllk-mlp*	GMM-UBM	17.6
Gabor+Δ	*ubmllk-mlp*	i-vector	11.5
MFCC	–	i-vector	11.3

Table 3. Gabor and MFCC feature results on NIST SRE08 male telephone data.

The GMM-UBM results demonstrated the applicability of the Gabor features on SRE08 data, though the results were suspected to improve with greater amounts of i-vector development data. The i-vector with LDA and WCCN approach to Gabor feature-based speaker recognition produced a 35% relative improvement over the baseline GMM-UBM system, using the *ubmllk-mlp* MLP training approach (11.5% vs. 17.6% EER). Furthermore, the Gabor features

had roughly the same performance as the MFCC features using the i-vector approach (11.5% vs. 11.3% EER), suggesting that the Gabor features are comparable to MFCC features in terms of performance on male telephone conversational data. Note that, due to the smaller number of GMM mixtures and small data size, only 100 i-vector dimensions were used, and 50 dimensions were kept after LDA processing.

We also investigated the score-level combination of Gabor+Δ features with the MFCC features on the ROSSI dataset. Combination was performed using an MLP with 2 hidden nodes and 1 hidden layer, implemented using Lnknet [16]. The EERs represent averaged EER values over 100 splits amongst the trials, where each split contained training and testing sub-splits. For each of the 100 splits, MLP weights were trained using the training sub-split, and EERs for each split were obtained by applying the MLP weights on the testing sub-split. The subsampling was performed even if there was only one system used, so that the standalone results would be consistent with the combination results. The score-level combination results are shown in table 4, where the *ubmllk-mlp* MLP training approach was used for Gabor feature extraction. The GMM-SVM approach was used for both the MFCC and Gabor+Δ systems.

System	EER (%)
MFCC	7.6
Gabor+Δ	7.4
MFCC + Gabor+Δ	7.0

Table 4. Score-level combination results for Gabor- and MFCC-based speaker recognition systems using the GMM-SVM approach and ROSSI database, with 100-split subsampling

According to table 4, the Gabor features were effective in score-level combination with MFCC features. Combining the MFCC and Gabor systems at the score-level gave an 8% relative EER improvement over the MFCC system standalone (7.0% vs. 7.6% EER), and a 5% relative EER improvement over the Gabor+Δ system standalone. While the improvements were not significant, they nevertheless suggested that Gabor features provided complementary information to the MFCC features in score-level combination, and that score-level combination of the two features was superior to feature-level combination in terms of EER.

5. DISCUSSION AND SUMMARY

This work demonstrates the applicability of 2D Gabor features for speaker recognition, and is the first known attempt to apply them to a large dataset such as the NIST speaker recognition framework. Using data with channel and environmental noise, we have demonstrated that Gabor features could perform as well as MFCC features standalone, and provide complementary information to MFCC features for score-level combination. Because of the large number of parameters involved in Gabor feature extraction, future gains could potentially be achieved through better dimensionality reduction approaches, and adjustments to the spectral and temporal modulation frequencies of Gabor features (the frequencies had been initially tuned for ASR). Future work could also reduce the computational costs of Gabor feature extraction, so that the features could be applied to larger NIST datasets.

6. ACKNOWLEDGEMENTS

This work is sponsored by Air Force Research Laboratory under contract FA8750-10-C-0214.

7. REFERENCES

[1] Battles, B. and Lawson, A. “NoTel: A Large, Naturally Noisy, Multi- Device Telephony Database For Speech And Speaker Recognition”, Under Review.

[2] Mesgarani, N., Stephen, D., and Shamma, S., “Representation of Phonemes in Primary Auditory Cortex: How the Brain Analyzes Speech”, in Proc. of ICASSP, 2007.

[3] Qiu, A., Schreiner, C., and Escabi, M., “Gabor Analysis of Auditory mid-brain Receptive Fields: Spectro-Temporal and Binaural Composition”, in Journal of Neurophysiology, vol. 90, pp. 456–476, 2003.

[4] Kleinschmidt, M and Gelbart, D., “Improving Word Accuracy with Gabor Feature Extraction”, in Proc. of Interspeech, 2002.

[5] Ravuri, S. and Morgan, N., “Using Spectro-Temporal Features to Improve AFE Feature Extraction for ASR”, in Proc. of Interspeech, 2010.

[6] Meyer, B.T., Ravuri, S., Schädler, M.R., and Morgan, N., “Comparing Different Flavors of Spectro-Temporal Features for ASR”, in Proc. of Interspeech, pp. 1269 – 1272, 2011.

[7] Mildner, V., Goetze, S., Mertins, A., and Kammeyer, K., “Optimization of Gabor Features for Text-Independent Speaker Identification”, in IEEE International Symposium on Circuits and Systems, 2007.

[8] Reynolds, D.A., Quatieri, T.F., and Dunn, R., “Speaker Verification using Adapted Gaussian Mixture Models”, in Digital Signal Processing, Vol. 10, pp 19–41, 2000.

[9] Campbell, W., Sturim, D., and Reynolds, D.,“Support Vector Machines using GMM Supervectors for Speaker Verification”, in IEEE Signal Processing Letters, Vol. 13, pp. 308 – 311, 2006.

[10] Schädler, M.R., Meyer, B.T., and Kollmeier, B., “Spectro-temporal Modulation Subspace-Spanning Filter Bank Features for Robust Automatic Speech Recognition”, submitted to J. Acoust. Soc. Am., 2011.

[11] Bonastre, J.F., Wils, F., and Meignier, S., “ALIZE, a free Toolkit for Speaker Recognition”, in ICASSP, Vol. 1, pp. 737–740 (2005).

[12] Joachims, T., “Making Large Scale SVM Learning Practical”, in Advances in kernel methods - support vector learning, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT-press, 1999.

[13] HMM Toolkit (HTK), http://htk.eng.cam.ac.uk

[14] Huijbregts, M., “Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled”, Ph. D Thesis, University of Twente, 2008

[15] Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., and Dumouchel, P., “Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification”, in Proc. of Interspeech, 2009.

[16] Lippmann, R.P., Kukolich, L.C., Singer, E., “LNKnet: Neural Network, Machine Learning, and Statistical Software for Pattern Classification, in Lincoln Laboratory Journal, Vol. 6, pp 249- 268, 1993.