Gabor Filterbank Features for Robust Speech Recognition

Ibrahim Missaoui[1] and Zied Lachiri[1,2]

[1] Laboratoire Signal, Images et Technologies de l'Information, École Nationale d'Ingénieurs de Tunis (ENIT), Université de Tunis El Manar, BP. 37, Le Belvedère, 1002, Tunis, Tunisia
[2] Département physique et instrumentation, Institut National des Sciences Appliquées et de Technologie (INSAT), Université de Carthage, BP. 676 centre urbain cedex, 1080 Tunis, Tunisia
{brahim.missaoui,zied.lachiri}@enit.rnu.tn

Abstract. Several research studies have shown that the robustness and performance of speech recognition systems can be improved using physiologically inspired filterbank based on Gabor filters. In this paper, we proposed a feature extraction method based on 59 two-dimensional Gabor filterbank. The use of these set of filters aims to extracting specific modulation frequencies and limiting the redundancy on feature level. The recognition performance of our feature extraction method is evaluated in isolated words extracted from TIMIT corpus. The obtained results demonstrate that the proposed extraction method gives better recognition rates to those obtained using the classic methods MFCC, PLP and LPC.

Keywords: Features extraction, Gabor filterbank, Robust speech recognition.

1 Introduction

Research on the selection of the best discriminate feature sets for Automatic Speech Recognition (ASR) system has been an area of great focus in various speech processing studies in last decades. The extraction and selection of these features significantly affects the performance of this system. Several of the proposed features are inspired and motivated by the speech processing strategies of the human auditory perception [9][12] e.g Linear Prediction coding (LPC) [15], Perceptual Linear Prediction (PLP) [4] and Mel-Frequency Cepstral Coefficients (MFCC) [1]. An extension of the PLP feature extraction known as RASTA PLP was later employed to suppress the temporal fluctuations noise [5]. These classic features are usually combined with the energy and with their first and second order derivation in order improve the recognition rate of ASR systems by incorporating information about the signal temporal dynamic.

Recently, 2-dimensional Gabor filters have been proposed by Kleinschmidt and Gelbart [7] in order to detect the spectro-temporal cues from sound signals. This was motivated by the physiological measurements findings showing that

A. Elmoataz et al. (Eds.): ICISP 2014, LNCS 8509, pp. 665–671, 2014.

the neurons of the primary auditory cortex of mammals are sensitive to particular patterns in the spectro-temporal representation. These spectro-temporal patterns are known as the spectro-temporal receptive fields (STRFs)[10][11].
In more recent work, the 2-D Gabor filters were exploited to improve an MFCC baseline by extracting spectro-temporal features which was used as direct input features for an HMM classifier [8][18][17]. In [14], the features are extracted from spectrograms derived from PNCCs (Power-Normalized Cepstral Coefficients) [6].

In this paper, we propose a new method based on Gabor filterbank for extracting relevant acoustic parameters. The used filterbank is composed of 59 2-D Gabor filters [18][13]. The adopted recognition system is based on Hidden Markov Model(HMM) using the HTK toolkit [19]. We validate our method by comparing the obtained recognition performance with those of the conventional techniques such MFCC, PLP and LPC.
This paper is organized as follows : after an introduction, Section 2 describes our acoustic parameters extraction method based on 2-D Gabor filterbank, the experiments and results are presented in Section 3. Section 4 provides a summary of our work.

2 Speech Recognition Based on Gabor Filterbank

2.1 The Proposed Feature Extraction Method

This section describes a feature extraction method based on Gabor filterbank for robust speech recognition. Our feature extraction method consists of five stages as depicted by a block diagram in the figure 1.

In the first stage, the power spectrum is calculated for each window segment obtained by applying a Short Time Fourier Transform (STFT) to the speech signal.

The second stage is the critical band analysis. It consists in warping the power spectrum along its frequency axis into the approximate Bark frequency. The result auditory spectrum is then convolved, to yield the critical-band power spectrum of human hearing, with a simulated critical-band masking curve [4].

Afterward, the outputs are weighted by an equal-loudness pre-emphasis in the third stage, which offers an approximation of the non equal sensitivity of human hearing.

The next stage in our feature extraction method is the Intensity-loudness power law. In this stage, the cubic-root amplitude compression is applied in order to simulate the power law of hearing.

In the last stage, the Gabor filterbank Auditory Spectrum features (GFAS features) are calculated by convolving the outputs with the Gabor filter bank consisting of 59 2-D Gabor filters [18] [13]. Each Gabor filter $g(n,k)$ is defined as the product of a complex sinusoid $s(n,k)$ and an Hanning envelope function $h(n,k)$ (with n is the time index and k is the frequency index) [18].

$$s(n,k) = \exp\left(i\,\omega_n(n-n_0) + i\,\omega_k(k-k_0)\right) \tag{1}$$

$$h(n,k) = 0.5 - 0.5\cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right)\cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right) \quad (2)$$

The terms ω_n, ω_k are the time and spectral modulation frequencies of the sinusoid $s(n,k)$ and, W_n and W_k are the time and frequency window lengths of the used window.

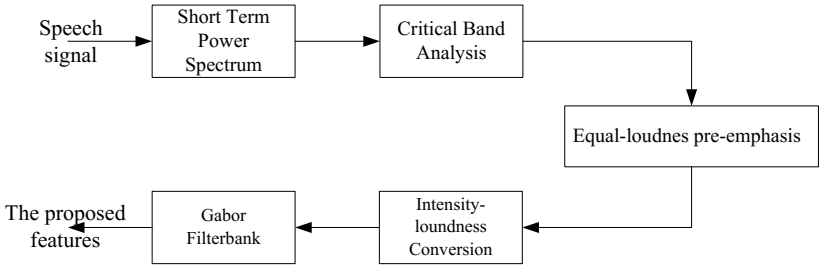

Fig. 1. Block diagram of the proposed feature extraction technique

The set of Gabor filters were chosen to exhibiting constant overlap and to covering a wide range of modulation frequencies, which offers an approximation of orthogonal filters, thereby allowing to limit the redundancy of output signal of the filter. The corresponding temporal and spectral modulation frequencies of the set of 59 Gabor filters are depicted in figure 2.

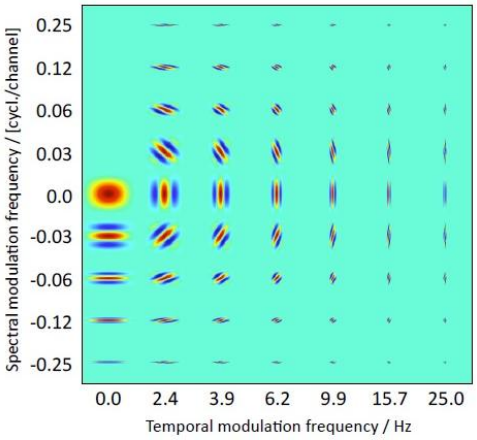

Fig. 2. Real components of the used 2-D Gabor filters

2.2 The HMM based Speech Recognition

The obtained GFAS features are then used as the input of HMM based speech recognition as shown in figure 3. The HMM (Hidden Markov models) are generative models based on doubly stochastic dynamical process characterized by an underlying stochastic process which is not observable [2][16]. They are composed of discrete stationary states that are connected by transitions. Each state generates over time a feature vectors o_t described by a probability distribution density. As illustrated in figure 4, each transition between state i and state j provides a sets of instantaneous probabilities distribution a_{ij}.

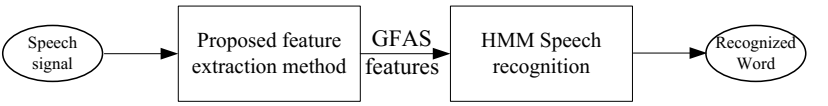

Fig. 3. General Scheme of a HMM based Speech Recognition System

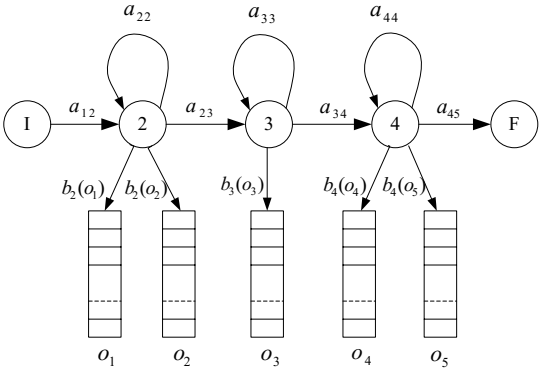

Fig. 4. A schematic of HMM with five states

Most HMM based speech recognition use continuous Gaussian mixture (GM) density to represent the output probabilities. The speech parameter vector $O = o_1, o_2, o_3, o_4, ..., o_T$ associated to each word is generated from the output probability distribution $b_j(o_t)$ which is computed as follows [16] :

$$b_j(o_t) = \sum_{k=1}^{K_j} c_{jk}\, N(o_t, \mu_{jk}, \vartheta_{jk}) \tag{3}$$

Where

$$N(o, \mu, \vartheta) = \frac{1}{((2\pi)^n |\vartheta|)^{\frac{1}{2}}}\ \exp\left(-\frac{1}{2}(o-\mu)^T \vartheta^{-1}(o-\mu)\right) \tag{4}$$

$N(o, \mu, \vartheta)$ is a multivariate Gaussian. c_{jk}, μ_{jk}, ϑ_{jk}, K_j and n is respectively the mixing weight, the mean vector, the covariance matrix and the number of mixture components and the dimensionality of o.

3 Experiments and Results

The performance of the proposed GFAS features was evaluated with 9240 isolated words speech and 3294 isolated-words used respectively for the learning phase and the recognition phase. These words were extracted from the TIMIT corpus [3]. This corpus consists of 630 speakers and there are 10 speech signal files with sampling frequency equal to 16 kHz for each speaker.

The HTK speech recognition toolkit [19] is exploited in the used Hidden Markov Models (HMM) based recognition system. Each isolated-word model of HMM topology consisted of five emitting states, each represented by a N Gaussian distribution mixtures with diagonal covariances and continuous density (HMM-N-GM). The value of N is chosen equal to 1, 2, 4 and 8 respectively. Table 1 summarizes the recognition rate (%) obtained by the proposed features, MFCC, PLP and LPC using HMM with N equal to 1, 2, 4 and 8 (HMM-1-GM, HMM-2-GM, HMM-4-GM and HMM-8-GM).

Table 1. The recognition rate of the proposed features, MFCC, PLP and LPC obtained using HMM-1-GM, HMM-2-GM, HMM-4-GM and HMM-8-GM

	Recognition rate with			
Technique	HMM-1-GM	HMM-2-GM	HMM-4-GM	HMM-8-GM
LPC	46.08	53.83	58.86	62.96
PLP	84.94	87.98	89.62	91.89
MFCC	84.76	88.49	90.26	92.96
Proposed features	89.65	94.69	96 .02	96.75

As illustrated in the table 1, we can observe that the recognition rate of the proposed the proposed GFAS features is performed better than the classic features MFCC, PLP and LPC in the different cases.

The highest percentage of the recognition rates is obtained using HMM-8-GM for the four features. For example, the recognition rate of the proposed features using HMM-8-GM is equal to 96.75, while the classic features MFCC, PLP and LPC had 92.96, 91.89 and 62.96 respectively. We can see also that LPC features is the worst performing in front of PLP, MFCC and GFAS.

4 Conclusion

In this paper, we presented a new extraction method of acoustic parameters for isolated-word speech recognition. The proposed method is based on 2-D Gabor filterbank. Their speech recognition performance is tested on isolated words extracted from the TIMIT database using Hidden Markov Models (HMM) with 1, 2, 4 or 8 Gaussian Mixture continuous densities.

The obtained results show that the proposed extraction method gives better recognition rates to those obtained using the classic methods such as MFCC, PLP and LPC.

References

1. Davis, S., Mermelstein, P. : Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing 28(4), 357–366 (1980)
2. Ephraim, Y., Merhav, N. : Hidden markov processes. IEEE Transactions on Information Theory 48(6), 1518–1569 (2002)
3. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. : DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N 93, 27403 (1993)
4. Hermansky, H. : Perceptual linear predictive (plp) analysis of speech. The Journal of the Acoustical Society of America 87, 1738 (1990)
5. Hermansky, H., Morgan, N. : Rasta processing of speech. IEEE Transactions on Speech and Audio Processing 2(4), 578–589 (1994)
6. Kim, C., Stern, R.M. : Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. In : Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 28–31 (2009)
7. Kleinschmidt, M., Gelbart, D. : Improving word accuracy with gabor feature extraction. In : Annual Conference of the International Speech Communication Association, INTERSPEECH (2002)
8. Lei, H., Meyer, B.T., Mirghafori, N. : Spectro-temporal gabor features for speaker recognition. In : IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4241–4244 (2012)
9. Lippmann, R.P. : Speech recognition by machines and humans. Speech Communication 22(1), 1–15 (1997)
10. Mesgarani, N., David, S., Shamma, S. : Representation of phonemes in primary auditory cortex : How the brain analyzes speech. In : IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. IV–765 (2007)
11. Mesgarani, N., Shamma, S. : Speech processing with a cortical representation of audio. In : IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5872–5875 (2011)
12. Meyer, B.T., Kollmeier, B. : Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Communication 53(5), 753–767 (2011)

13. Meyer, B.T., Ravuri, S.V., Schädler, M.R., Morgan, N. : Comparing Different Flavors of Spectro-Temporal Features for ASR. In : Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1269–1272 (2011)
14. Meyer, B.T., Spille, C., Kollmeier, B., Morgan, N. : Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition. In : Annual Conference of the International Speech Communication Association (INTERSPEECH), vol. 15, p. 20 (2012)
15. O'Shaughnessy, D. : Linear predictive coding. IEEE Potentials 7(1), 29–32 (1988)
16. Rabiner, L. : A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
17. Ravuri, S.V., Morgan, N. : Using spectro-temporal features to improve AFE feature extraction for ASR. In : Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1181–1184 (2010)
18. Schädler, M.R., Meyer, B.T., Kollmeier, B. : Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. The Journal of the Acoustical Society of America 131, 4134 (2012)
19. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P. : The HTK book (Revised for HTK version 3.4.1). Cambridge University Engineering Department (2009)