Morphological Processing of a Dynamic Compressive Gammachirp Filterbank for Automatic Speech Recognition

Joyner Cadore, Carmen Peláez-Moreno, and Ascensión Gallardo-Antolín

Universidad Carlos III de Madrid, Escuela Politécnica Superior,
Avda. de la Universidad 30, 28911 Madrid, Spain
{jcadore,carmen,gallardo}@tsc.uc3m.es
http://gpm.tsc.uc3m.es/

Abstract. The Dynamic Compressive Gammachirp ([8]) is presented for producing auditory-inspired feature extraction in Automatic Speech Recognition. The proposed acoustic features combine spectral subtraction and two-dimensional non-linear filtering technique most usually employed for image processing: morphological filtering. These features have been proven to be more robust to noisy speech than those based on simpler auditory filterbanks like the classical mel-scaled triangular filterbank, the Gammatone filterbank and the passive Gammachirp in a noisy Isolet database.

Keywords: spectral subtraction, morphological processing, automatic speech recognition, auditory-based features.

1 Introduction

Machine performance in Speech Recognition tasks is still far from that of humans.

Among non-satisfactorily tackled challenges is background noise. One way to address these limitations is trying to imitate human acoustic capabilities, e.g. finding a more suitable auditory model. In this paper we focus on the noise problem where humans are known to perform remarkably well whilst machines still lag behind [13].

There are many solutions inspired by the Human Auditory System (HAS) aimed at solving this issue, e.g. feature extraction based on the well-known mel-frequency cepstral coefficients (MFCC), and on the Gammatone-based coefficients (GTC) are some of many examples. Other solutions use the so-called *spectro-temporal features*, that consider both the time and frequency domains in the feature extraction stage ([10]).

Following that line of work, the authors presented a morphological filtering over a cochleogram (i.e. an auditory spectrogram) of a noisy signal to mimic some properties of the HAS, such as frequency and temporal masking [3, 4]. Two types of basic cochleograms were compared: the first, based on classical mel-scaled critical bands and the second, on gammatone auditory filters presenting comparable results. In this paper, we propose the use of more detailed formulation of the auditory filtering procedure based on successive improvements of the gammatone formulation by Irino and Patterson: the dynamic compressive gammachirp (dcGC) auditory filterbank [8].

Gammatones (GT) have been already employed in a number of papers for ASR ([15] among others) while passive gammachirp (pGC) of [8] is less widespread ([9]). However, to our knowledge, only preliminary experiments on syllable recognition have been deployed for dcGC with the purpose of showing the scale-shift covariance properties of this auditory model that provides a better way for adapting ASR acoustic models to vocal tract length variations ([12]). A different approach for introducing dynamic auditory filtering is based on the application of the Dyn non-linear operator applied on the compressive gammachirp (cGC) [5] demonstrating robustness improvements over the conventional MFCC in several noisy conditions on TIMIT.

In this paper we have found that dcGC outperforms cGC in noisy mismatched conditions specially in combination with the well-known Spectral Subtraction (SS) preprocessing method. The morphological postfiltering applied to the resulting cochleogram also provides enhanced performance (though marginally). The new resulting features have been tried on a hybrid MLP/HMM recognizer.

This paper is organized as follows: in section 2 we describe the alternative methods for auditory filtering aforementioned. Section 3 presents the pre and postprocessing stages of our feature extraction method. Section 4 describes the experiments and results obtained to end with the conclusions and futher work in section 5.

2 Gammachirp auditory filters

The Gammachirp auditory filter [8] is an extension of the Gammatone filter. The impulse response of a gammachirp filter is defined by:

$$g(t) = kt^{n-1}exp(-2\pi b\mathrm{ERB}(f_c)t) \times exp(j2\pi f_c t + jc\ln(t) + j\phi) \tag{1}$$

where n is the order of the filter, k defines the output gain, b defines the envelope of the gamma distribution, c is the chirp factor, f_c is the filter's central frequency, ϕ is the phase and ERB is the Equivalent Rectangular Bandwidth defined in [11]. When $c = 0$, eq. (1) reduces to the impulse response of the Gammatone filter.

The Fourier magnitude spectrum of the Gammachirp filter is:

$$\left|G_C(f)\right| = A \cdot \left|G_T(f)\right| \cdot exp(c\theta(f)) \tag{2}$$

$$\theta(f) = arctan(\frac{f - f_c}{b\mathrm{ERB}(f_c)}) \tag{3}$$

where $|G_T(f)|$ is the Fourier magnitude spectrum of the Gammatone filter.

From (2), (3) it is possible to obtain the next three types of Gammachirp filters [8]:

- The Passive Gammachirp (pGC): level-independent and representing the passive basilar membrane.
- The Compressive Gammachirp (cGC): level-dependent and simulating the active mechanism in the cochlea.
- The Dynamic Compressive Gammachirp (dcGC): including a fast-acting level control circuit for the cGC filter, two-tone suppression and compression.

3 Spectral Subtraction and Morphological Filtering

The filterbanks described in section 2, furthermore the very well-known Mel-scaled triangular filterbank and Gammatone filterbank, have been embedded in a feature extraction system that includes spectral subtraction as a preprocessing stage and a morphological filtering as postprocessing aimed at imitating both temporal and instantaneous masking in the HAS.

3.1 Spectral Subtraction

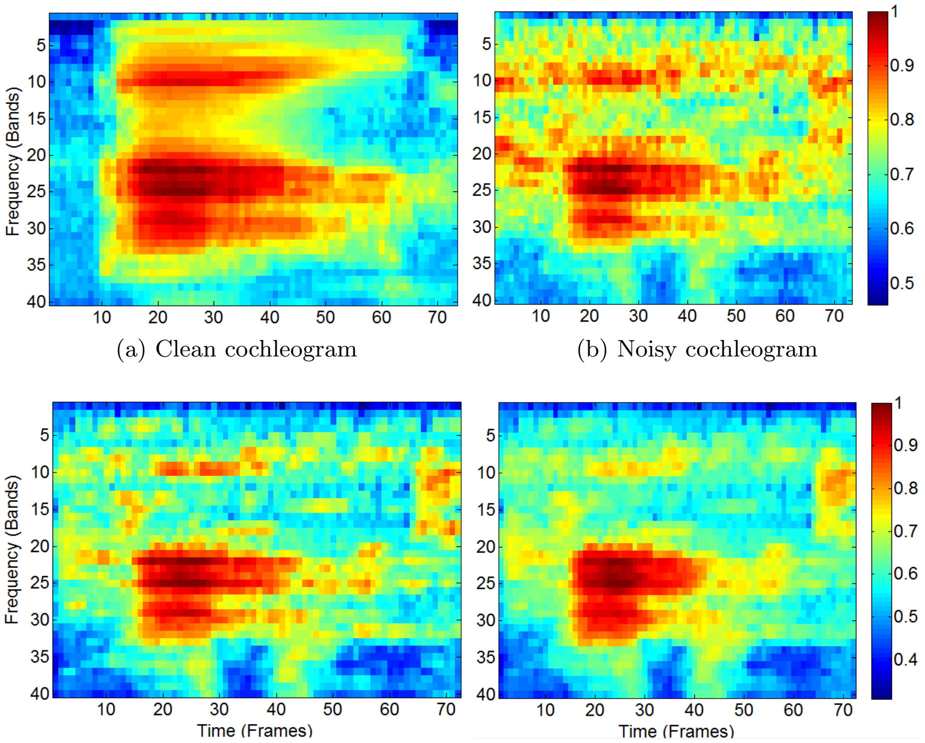
Spectral Subtraction (SS) is a classical procedure to remove noise from speech [1]. Figures 1a and 1b shows the cochleogram of a clean and noisy speech sample, respectively. Spectral subtraction obtains an estimate of the density spectrum of the noise and performs a subtraction in the frequency domain. The SS was applied in the magnitude domain. Fig. 1c depicts the resulting cochleogram of the so cleaned signal.

3.2 Morphological Filtering of cochleograms

Auditory masking has been largely studied as regarding the influence of some frequencies on others simultaneously present in the spectrum, or *simultaneous masking*, or as regarding the influence of the same frequencies at different time instants, or *temporal masking*. Therefore its effects can be observed both in time and frequency domains requiring a two dimensional representation to jointly consider the two of them.

The application of an auditory motivated filterbank (as those described in section 2) produces a more uniform representation of the simultaneous masking effects that is certainly more amenable for a computational modelling since the asymmetrical masking threshold becomes *almost* independent of the scaled frequency. An auditory spectrogram (sometimes referred as *cochleogram*) substitutes the usual linear spectral representation by auditory motivated filterbanks uniformly distributed in a scaled frequency.

On the other hand, Mathematical Morphology (MM) is a theory for the analysis of spatial structures [14] whose main application domain is in image processing as a tool for thinning, pruning, structure enhancement, object marking, segmentation and noise filtering. Thus, Morphological Filtering (MF) becomes an adequate operation to consider both domains of auditory masking.



(a) Clean cochleogram

(b) Noisy cochleogram

(c) Noisy cochleogram + SS

(d) Noisy cochleogram + SS + MF

Fig. 1: Resulting auditory cochleograms of a utterance with added babble noise at 5dB.

3.3 Morphological operations

We use S with implicit frequency band index n and temporal frame k to represent a cochleogram $S(n, k)$. Then erosion and dilation with a given mask M can be represented as matrix operations:

$$S \ominus M = \{p \in \mathbb{R}^2 \mid p = m - s, m \in M,\ \ s \in S\} \tag{4}$$

$$S \oplus M = \{p \in \mathbb{R}^2 \mid p = s + m, s \in S, m \in M\} \tag{5}$$

Erosion is used to shrink or reduce objects, while dilation, being the dual to erosion, produces an enlargement. Both are irreversible.

Opening and closing are used to remove small objects in images, typically noise, their behaviour with respect to, for instance, salt and pepper noise, being dual to each other. The composite opening and closing operation adopt the following form:

$$S \circ M = (S \ominus M) \oplus M \tag{6}$$

$$S \bullet M = (S \oplus M) \ominus M \tag{7}$$

We use the opening operator over the cochleogram with the discrete representation of the mask depicted in fig. 3 to obtain:

$$S'' = S' + S' \circ M \tag{8}$$

The resulting cochleogram S'' of a sample noisy signal is shown in fig. 1d. Finally an DCT is performed to obtain $\hat{S}$ as shown in fig. 2.

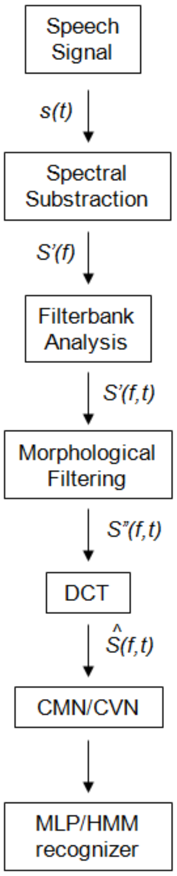


Fig. 2: Block diagram of the ASR system.

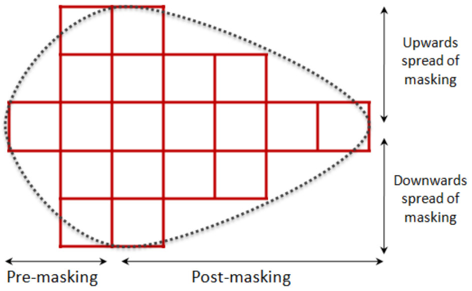


Fig. 3: Auditory-motivated mask (dotted line) and its discretization (solid line). The horizontal and the vertical axis represents time and frequency scales, respectively.

4 Experiments

4.1 General Description and Feature Extraction

The block diagram of the ASR system used in the experimentation is depicted in fig. 2.

Four different types of auditory filterbanks are considered: triangular mel-scaled, Gammatone, passive Gammachirp and dynamic compressive Gammachirp filters, yielding Mel-Frequency Cepstrum Coefficients (MFCC), Gammatone-based (GTC), passive Gammachirp-based (pGC) and dynamic compressive Gammachirp-based (dcGC) features, respectively.

In all the cases, speech is analyzed using a 25 ms window every 10 ms. The corresponding auditory filterbank is composed of 40 bands. After the DCT, coefficients C_0 to C_{12} are kept. Adding their corresponding delta and acceleration coefficients compose 39 dimensional vectors. The last step in the feature extraction stage was applying mean and variance normalization on either type of coefficient.

4.2 ISOLET Testbed

In the experimentation, we use the ISOLET testbed [7]. ISOLET is a database of letters of the English alphabet spoken in isolation. The database consists of 7800 spoken letters (two productions of each letter pronounced by 150 different speakers). Specifically, we use a version called Noisy-ISOLET: the speech signals of ISOLET plus 8 different noise types at different SNRs (clean, 0dB, 5dB, 10dB, 15dB and 20dB).

The experiments using the ISOLET testbed are performed over an hybrid MLP/HMM ASR system [2]. A context of 5 frames is used yielding an input of 195 elements to the MLP. The hybrid MLP/HMM system is tested in two different conditions: *mismatched*, where the system is trained using clean speech and *matched* where the training set is composed of a balanced combination of speech contaminated with the different noises of the database at several SNRs. A 5-fold cross-correlation procedure has been employed to improve statistical significance [6].

4.3 Results

Table 1 summarizes the experiments performed to study the impact of the Spectral Subtraction (SS), the Auditory filtering and the Morphological Filtering (MF).

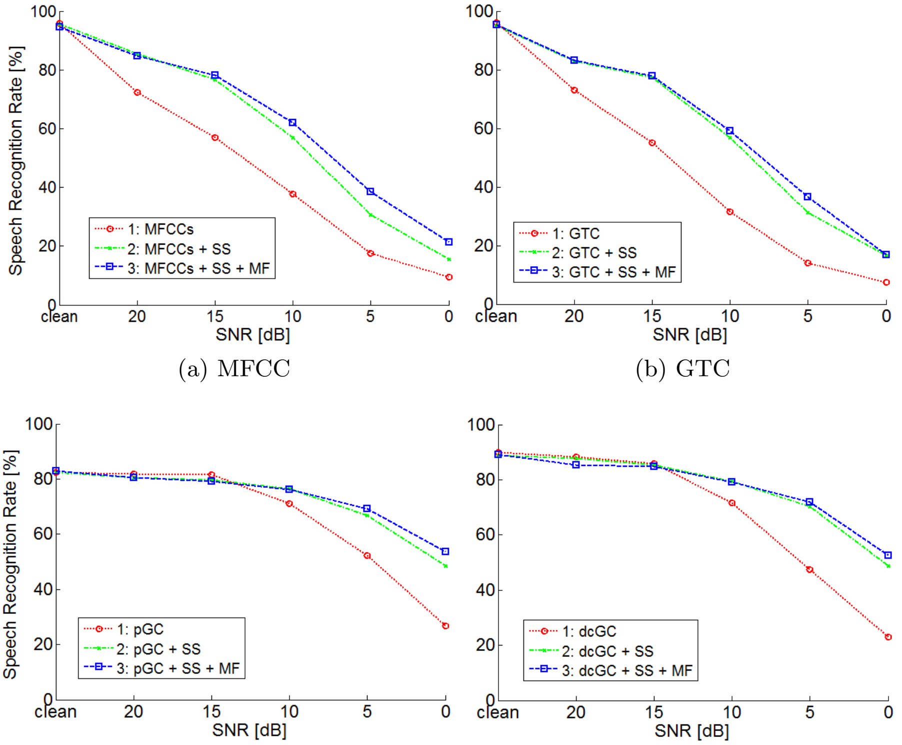
Table 1: Recognition results in terms of *WER* [%] and 95% confidence intervals.

Features	Mismatched	Matched
MFCC	51.80 ± 1.24	16.45 ± 0.92
MFCC + SS	40.85 ± 1.22	16.95 ± 0.93
MFCC + SS + MF	37.03 ± 1.20	17.05 ± 0.93
GTC	53.78 ± 1.24	17.15 ± 0.94
GTC + SS	40.28 ± 1.22	16.95 ± 0.93
GTC + SS + MF	38.50 ± 1.21	16.85 ± 0.93
pGC	34.03 ± 1.18	25.78 ± 1.09
pGC + SS	27.60 ± 1.11	26.63 ± 1.10
pGC + SS + MF	26.45 ± 1.09	27.95 ± 1.11
dcGC	32.40 ± 1.16	20.08 ± 0.99
dcGC + SS	23.25 ± 1.05	20.38 ± 1.00
dcGC + SS + MF	22.98 ± 1.04	20.18 ± 1.00

As for the *mismatched* condition when the auditory filterbank is considered alone, pGC and dcGC features achieve better results than MFCC and GTC, being the differences statistically significant. This fact shows the robustness of pGC and dcGC parameters in noisy conditions. The use of SS clearly improves the corresponding baselines where the best results are obtained with the dcGC + SS features, being the differences statistically significant again with respect. The sequential use of both techniques (SS + MF) improves the recognition rates of the system compared to the baseline and SS cases, specially for the MFCC parameterization. Nevertheless, dcGC + SS + MF achieves the best performance with a relative error reduction of around 38% with respect to MFCC + SS + MF.

For the *matched* condition, best results are obtained with MFCC and GTC features. Besides, since no significant improvements are achieved by using SS or SS + MF the aforementioned techniques seem to be more suitable for the *mismatched* case.

Figure 4 shows the Recognition Rates achieved by the different techniques as a function of the SNR for the *mismatched* condition. As can be observed,



(a) MFCC

(b) GTC

(c) pGC

(d) dcGC

Fig. 4: SRR [%] vs SNR. Mismatched case.

the behaviour of MFCC and GTC features is rather different to the one shown by pGC and dcGC. The performance of the two first types of features degrades significantly for all SNRs, whereas the recognition rates achieved by pGC and dcGC only decrease drastically for the lower SNRs (0 dB - 10 dB). The combination of SS or SS + MF techniques with MFCC and GTC features improves the recognition rates for all SNRs. However, for pGC and dcGC, SS or SS + MF is only effective for SNRs below 15 dB. In any case, although in clean conditions, MFCC and GTC (alone or with SS or SS + MF) report the best results, the corresponding systems based on pGC and dcGC features achieve similar results for high SNRs and clearly outperform MFCC and GTC for low SNRs. Finally, when comparing pGC and dcGC, it can be observed that both sets of features exhibit similar trends, but dcGC is better than pGC for clean and high SNRs conditions.

5 Conclusions and future work

The Dynamic Compressive Gammachirp have been employed for producing auditory-inspired feature extraction in Automatic Speech Recognition. The proposed acoustic features also combine spectral subtraction and two-dimensional non-linear filtering technique most usually employed for image processing: morphological filtering. These features have been proven to be more robust to noisy speech than those based on simpler auditory filterbanks like the classical mel-scaled triangular filterbank, the Gammatone filterbank and the passive Gammachirp in a noisy Isolet database.

Future lines of work include testing different types of masks and morphological operations and further study the synergies between the effects of the mask and the dcGC-based features.

Acknowledgments. This work has been partially supported by the Spanish Ministry of Science and Innovation CICYT Projects No. TEC2008-06382/TEC and No. TEC2011-26807. We would like to thank Prof. Toshio Irino for kindly providing the software for Gammachirp auditory filters.

References

1. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79. vol. 4, pp. 208–211. IEEE (1979)
2. Bourlard, H., Morgan, N.: Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. Adaptive Processing of Sequences and Data Structures pp. 389–417 (1998)
3. Cadore, J., Gallardo-Antolín, A., Peláez-Moreno, C.: Morphological processing of spectrograms for speech enhancement. Advances in Nonlinear Speech Processing pp. 224–231 (2011)
4. Cadore, J., Valverde-Albacete, F.J., Gallardo-Antolín, A., Peláez-Moreno, C.: Auditory-inspired morphological processing of speech spectrograms. Cognitive Computation (Jul 2012 (accepted))
5. Gauci, O., Debono, C., Micallef, P.: A nonlinear feature extraction method for phoneme recognition. In: Electrotechnical Conference, 2008. MELECON 2008. The 14th IEEE Mediterranean. pp. 811–815 (2008)
6. Geisser, S.: The predictive sample reuse method with applications. Journal of the American Statistical Association 70, 320–328 (1975)
7. Gelbart, D., W., H., Holmberg, M., Morgan, N.: Noisy ISOLET and ISOLET testbeds (Feb 2011), `http://www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet/`
8. Irino, T., Patterson, R.D.: A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE Transactions on Audio, Speech, and Language Processing 14(6), 2222–2232 (Oct 2006)
9. Maganti, H., Matassoni, M.: A Level-Dependent Auditory Filter-Bank for Speech Recognition in Reverberant Environments. Twelfth Annual Conference of the International Speech Communication Association (2011)

10. Meyer, B., Kollmeier, B.: Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Communication (53), 753–767 (2010)
11. Moore, B., Glasberg, B.: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The Journal of the Acoustical Society of America 74, 750 (1983)
12. Patterson, R., Walters, T., Monaghan, J., Feldbauer, C., Irino, T.: Auditory speech processing for scale-shift covariance and its evaluation in automatic speech recognition. Audio, Transactions of the IRE Professional Group on pp. 3813–3816 (Dec 2009)
13. Peláez-Moreno, C., García-Moral, A., Valverde-Albacete, F.: Analyzing phonetic confusions using formal concept analysis. The Journal of the Acoustical Society of America 128(3), 1377–1390 (2010)
14. Serra, J., Soille, P. (eds.): Mathematical Morphology and its Application to Image Processing. Computational Imaging and Vision, Kluwer Academic (1994)
15. Yin, H., Hohmann, V., Nadeu, C.: Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency. Speech Communication (53), 707–715 (2010)