Monaural Speech Separation Based on a 2D Processing and Harmonic Analysis

Azam Rabiee[1,4], *Saeed Setayeshi*[2], *Soo-Young Lee*[3,4]

[1] Department of Computer, Islamic Azad University, Dolatabad Branch, Isfahan, Iran
[2] Faculty of Nuclear Engineering and Physics, Amirkabir University of Technology, Iran
[3] Department of Electrical Engineering, Korea Advanced Institute of Science & Technology, Korea
[4] Brain Science Research Center, Korea Advanced Institute of Science & Technology, Korea

azrabiee@gmail.com, setayesh@aut.ac.ir, sylee@kaist.ac.kr

Abstract

This paper proposes a new Computational Auditory Scene Analysis (CASA) approach based on a 2D spectro-temporal analysis and harmonic separation. The 2D processing, so-called Grating Compression Transform (GCT), analyzes the spectro-temporal content of the spectrogram, mimicking the processing of the primary auditory cortex. The estimated pitches from the GCT analysis are used for separation using harmonic magnitude suppression (HMS). A powerful aspect of our model is requiring no prior training on a specific training corpus. A baseline system based on the harmonic separation is designed for comparison. Since the baseline system is similar to the proposed except the auditory-cortex-like analysis, the SIR results illustrate its importance in this task.

Index Terms: monaural speech separation, computational auditory scene analysis, spectro-temporal analysis, harmonic magnitude suppression

1. Introduction

Computational Auditory Scene Analysis (CASA) is addressed to model the human auditory system in monaural speech separation. The role of the auditory peripheral analysis is widely utilized to transfer the waveform signal into 2D spectrogram. In contrast, a few papers explored the role of cortical mechanisms in organizing complex auditory scenes [1]. Going a step forward, this paper proposes a new CASA, investigating the role of the primary auditory cortex. Specifically, auditory cortex in or near Heschl's gyrus as well as in the planum temporale are involved in sound segregation [2]. One of the essential cues for monaural speech separation is pitch. Recently, an auditory-cortex-like spectro-temporal analysis, called GCT, has been presented in [3] for multi-pitch extraction. In GCT, localized time-frequency regions of the spectrogram are analyzed to extract pitch candidates. This paper proposes the CASA based on the spectro-temporal GCT analysis followed by harmonic magnitude suppression (HMS) for separation. The harmonic separation is motivated from Shamma's researches [1][4] which have used assumption-free harmonic templates based on a biologically plausible mechanism for periodicity pitch detection. The details of the proposed model are presented in Section 2. Section 3 provides the experimental results, and finally the conclusion is in section 4.

2. Model Overview

The schematic diagram of the proposed CASA system based on the spectro-temporal GCT analysis and HMS is demonstrated in Figure 1. The first stage is a peripheral analysis following the procedure in human's early auditory system which converts the 1D mixture input signal into a 2D auditory spectrogram [5]. Meanwhile, we estimates pitch values per spectrogram frame based on the 2D processing of GCT, inspired from the auditory cortex function. Due to a limitation of the GCT analysis, in the current system, GCT is applied on a linear-frequency spectrogram rather than the auditory log-frequency spectrogram. Every multi-pitch extraction approaches suffer from two sources with closed pitch values. Utilizing the GCT in the current system has two benefits in this case: (1) The GCT considers the fact that the pitch gap of two sources in high frequencies is greater than the low ones. Regarding its inherent localized process, it generates candidates for each localized patch. (2) The GCT maintains the separability of two pitch trajectories with equal pitch values but different pitch derivatives, due to its explicit representation of the underlying temporal trajectories [3]. In the next stage, multi-pitch selection generates estimated pitch lines from the pitch candidates of the previous stage. Finally, in the separation stage, harmonic patterns corresponding to the estimated pitch lines are generated using harmonic templates. Separation is carried out in the auditory spectrogram domain and the output segregated spectrograms can be analyzed further for recognition or any other applications. Details of the proposed model are explained in the following subsections.

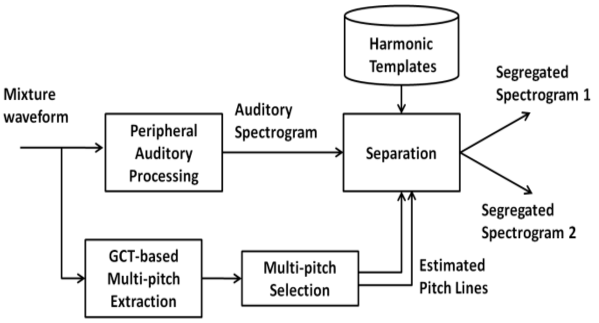

Figure 1. *Schematic diagram of the proposed CASA*

2.1. Auditory Spectrogram

In brief, the peripheral analysis consists cochlear filter bank, hair cell transduction, lateral inhibitory network, and midbrain integration [1] [5]. The cochlear filter bank contains a bank of 128 overlapping band pass filters with center frequencies uniformly distributed along a logarithmic frequency axis (x), over 5.3 oct (24 filters/octave). The impulse response of each filter is denoted by $f(t;x)$. These cochlear filter outputs $y_{coch}(t, x)$ are transduced into auditory-nerve patterns $y_{AN}(t, x)$ by a hair cell stage consisting of a high pass filter, a nonlinear compression $g(\cdot)$, and a membrane leakage low-pass filter $w(t)$ accounting for decrease of phase-locking on the auditory nerve beyond 2 kHz. The final transformation simulates the action of a lateral inhibitory network. The LIN is simply approximated by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectifier to produce $y_{LIN}(t,x)$. The final output of this stage is obtained by integrating $y_{LIN}(t,x)$ over a short window, $\mu(t,\tau)=e^{-t/\tau}u(t)$, with time constant $\tau=8$ ms mimicking the further loss of phase locking observed in the midbrain. Given a discrete-time signal *s(t)*, the mathematical formulation for these stages can be summarized as follows:

$$y_{coch}(t,x)=s(t)*_t f(t;x) \tag{1}$$

$$y_{AN}(t,x)=g\left(\partial_t y_{coch}(t,x)\right)*_t \omega(t) \tag{2}$$

$$y_{LIN}(t,x)=\max(\partial_x y_{AN}(t,x),0) \tag{3}$$

$$p(t,x)=y_{LIN}(t,x)*_t \mu(t,\tau) \tag{4}$$

where $*_t$ is convolution in time domain, and $p(t,x)$ is the spectrogram. For detailed information, refer to [5].

2.2. GCT-based Multi-pitch Extraction

In GCT, a 2D Fourier transform is computed over a localized time-frequency regions of a narrow band spectrogram. Similar to [3], the process begins with 512-point STFT using a 25 ms Hamming window with 1 ms frame shift. At every grid point (i,j) of the output, a patch $P_{ij}(f,t)$ of size 50 ms by 700 Hz is extracted. The patches have shifts of 5 ms in time and 100 Hz in frequency. Additionally, the patch mean value is subtracted before any further process. The GCT analysis on each patch is carried out by a 2D Gaussian window and a 512 x 256-point 2D FFT. The analysis compressed the harmonic patterns of each spectrogram patch to concentrated region in GCT plane. Then the vertical distance of the peaks from the GCT plane center is utilized to calculate the pitch candidates similar the approach in [3]. We have used the approach to extract pitch candidates and then a multi-pitch selection mechanism chooses the best pitch among the candidates per each frame, in the next stage.

2.3. Multi-pitch Selection

One of the drawbacks of the GCT analysis is generating many candidates per time frame. Hence, a multi-pitch selection or tracking mechanism is required to select the best estimation among the candidates. Before any further process, it is necessary to prune false candidates. False candidates are doubling/halving candidates, and the candidates belong to non-harmonic or silence patches. Every pitch extraction technique suffers from the doubling/halving phenomena. Here, we simply skip the pitch if its value is double or half times of another. To prune the false candidates obtained from non-harmonic patches, the harmonic templates, so-called *six-first harmonics without high frequency* (Figure 2(d)), are utilized. $\mathrm{w}(\mathrm{t}_0,\mathrm{F}_0)$, a weight value for each F_0 candidate at time frame t_0 is calculated based on the degree of the match between the auditory spectrogram $\mathrm{p}(\mathrm{t}_0,\mathrm{x})$ at time frame t_0, and the harmonic template $\mathrm{T}(\mathrm{x};\mathrm{F}_0)$, as follows:

$$w(t_0,F_0)=\frac{N}{M}\sum_{x=1}^{M}p(t_0,x)T(x;F_0) \tag{5}$$

where N is number of F_0 candidates in time frame t_0, M is number of frequency channels in the auditory spectrogram and x is the frequency index. Hence, the false pitch candidate is the one with low weight. However, a silence frame is defined based on its energy in auditory spectrogram. For comparison, we used two different multi-pitch selectors on the pruned candidates:

(i) Median-based selection

A simple k-mean clustering is applied on whole the pruned candidates. The estimated pitch value in each cluster per frame is the median one. To keep the time continuity of the estimated pitch values, the candidates of the previous and next frames are also added to the current candidates for median selection, similar to [3].

(ii) Hierarchical clustering + Meidan-based selection

The process begins with segmentation due to the silence frames. After pruning the short-length segments, the k-mean clustering is carried out on each segment. The estimated pitch values per frame are the median of each segment-based cluster. Now, each segment has two pitch trajectories. For assigning the trajectories to the speakers a higher level k-mean is carries out on the trajectory centers which are calculated by the average of their pitch values. Finally, a refinement is required to correct the time-overlapped trajectories in each cluster. Since some segments are single source, both trajectories may assign to one speaker. In this case, the sub-cluster centers should be close enough. Hence, a refinement is performed to merge both clusters of the segment and median selection gives the estimated pitch values. For other time-overlapped trajectories, a simple distance-based rule is utilized to avoid same cluster for both.

Finally, the estimated pitch values are mapped into the auditory spectrogram domain, and are utilized in separation using the harmonic templates.

2.4. Harmonic Templates and Separation

The harmonic templates are calculated following the approach in [4] from the auditory spectrograms of the sentences in TIMIT database. They correspond to the pitches from 90 Hz to 300 Hz. The original harmonic template for $F_0=125\,Hz$ is shown in Figure 2(a), as an example. The templates do not utilize any assumption on the number of pitches and allow us to extract harmonic structures at any given time instant [1]. One may use any templates including binary (1/0) mask. The key thing is getting some templates that can be used as a "sieve" to pull out the harmonic profiles. For comparison, two more sets of harmonic templates are generated from the original harmonic templates. As shown in Figure 2(b), the second set utilizes 1 as the value of each peak and high frequencies. For the third set, shown in Figure 2(c), the value for high frequencies is obtained from the power ratio of the extracted signals in low frequencies.

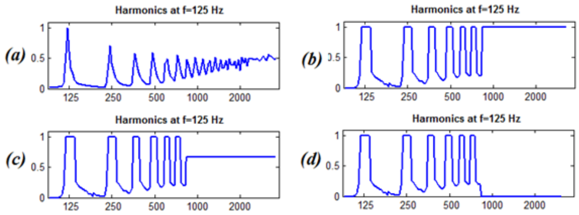

Figure 2. *Harmonic templates for F_0=125Hz. x axis is frequency (Hz) (a) original. (b) six-first harmonics with fixed high frequency. (c) six-first harmonics with weighted high frequency. (d) six-first harmonics without high frequency.*

Finally, the estimated pitch lines are utilized in separation using a point by point multiplication of the auditory spectrogram and the corresponding harmonic templates along the frequency axis.

3. Experimental Results

3.1. Baseline Model

A monaural speech separation based on harmonic analysis is designed as a baseline-CASA system for comparison. Figure 3 shows the schematic diagram of the baseline model. Basically, the baseline and the proposed models are same in the peripheral auditory processing and HMS for separation, but different in the multi-pitch extraction approaches. Instead of the auditory-cortex-like spectro-temporal analysis, the baseline system utilizes a harmonic-based multi-pitch extraction. The multi-pitch extraction is basically a comb-filter-based method which is used in the state-of-the-art single channel speech separation algorithm [6], similarly. The multi-pitch extraction is developed using the harmonic templates. Then, a multi-pitch tracking rule is used based on the slow transition of the fundamental frequency to generate the pitch trajectories. Finally, a clustering is performed to divide the pitch trajectories into clusters. The details are described in the following.

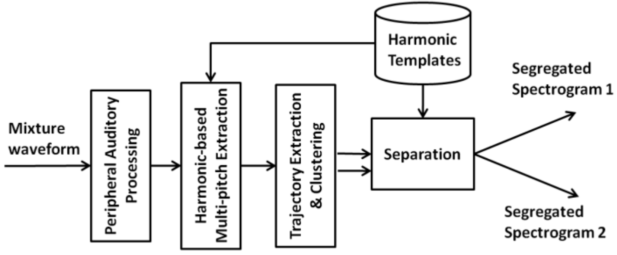
Mixture waveform
Peripheral Auditory Processing
Harmonic-based Multi-pitch Extraction
Trajectory Extraction & Clustering
Harmonic Templates
Separation
Segregated Spectrogram 1
Segregated Spectrogram 2

Figure 3. *Schematic diagram of the baseline CASA*

3.1.1. Harmonic-based Multi-pitch Extraction

In the baseline-CASA system, a multi-pitch extraction is developed based on harmonic analysis motivated from [1]. This harmonic analysis can convert the time-frequency spectrogram into a 3D time-frequency-pitch representation, called *harmonic patterns.* The harmonic patterns are extracted by

$$p_h(t_0, x, h) = p(t_0, x) \cdot_x T(x; h)\ . \tag{6}$$

Here $T(x; h)$ is the *six-first harmonics without high frequency* harmonic templates shown in Figure 2(d) and the subscript '$\cdot_x$' denotes the point-by-point multiplication along the frequency axis x. Hence, the harmonic patterns contain sieved spectrogram in different fundamental frequencies, h. These harmonic profiles are utilized to extract pitch tracks. In fact, the pitch candidates correspond to the harmonic patterns that are most similar to the original auditory spectrogram or carrying higher energy. The doubling/halving candidates and the candidates belong to the silence frames are pruned similar the proposed method.

3.1.2. Trajectory Extraction and Clustering

Pruned pitch values are passed into a multi-pitch tracker to form smooth pitch track and to extract pitch trajectories. A pitch trajectory is a sequence of neighbor pitch candidates which can be related to one segment in every clean speech signal. A plausible rule to track the pitch changes frame-by-frame is used, similar to [5][7]: Maximum pitch changes in two consequent frames should be less than 5% of the preceding pitch, while the frame shift is 8 ms. The short trajectories are simply removed. Now, the trajectories should be divided into two clusters. We defined an *average weighted pitch* value for each trajectory. The value for the trajectory $T_r = \{h_{t_0}, h_{t_0+1}, \dots, h_{t_{0+L_r-1}}\}$ is defined by

$$A(T_r) = \frac{1}{L_r} \sum_{t=t_0}^{t_0+L_r-1} h_t\, S(t, h_t) \tag{7}$$

where h_t is the pitch candidate in the frame t; and $S(t, h_t)$ is the energy of the corresponding harmonic pattern, and can be regarded as the strength or the weight of the candidate. A k-mean clustering and refinement are carried out on the average weighted pitch of each trajectory, similar to the proposed method.

3.2. Separation Results

Two clean speeches from male and female have been added instantaneously in time and are feed into the proposed system. Figure 4 shows the spectrograms of the segregation. Both left figures show the spectrogram of clean signals. The middle one is the mixture spectrogram and the two rights are the segregated spectrograms. To measure the performance of the segregation, correlation coefficients are computed between the original and recovered spectrograms, similar [1]. ρ_{Base} is the correlation coefficient between the two clean signals. $\rho_{Seg\,1(or\ 2)}$ and $\rho_{Conf\,1(or\ 2)}$ measure the similarity between the original and segregated speeches of the same speakers and the crossing speakers, respectively. We quantified the performance of the system by computing the correlation coefficients from 400 speech mixtures of both genders in TIMIT database. Figure 5 shows the histogram of ρ_{Base} (base), ρ_{Seg} (segregated) and ρ_{Conf} (confusion). The results show ρ_{Seg} values are significantly higher than ρ_{Conf}. However, the ρ_{Conf} values are almost around the ρ_{Base}.

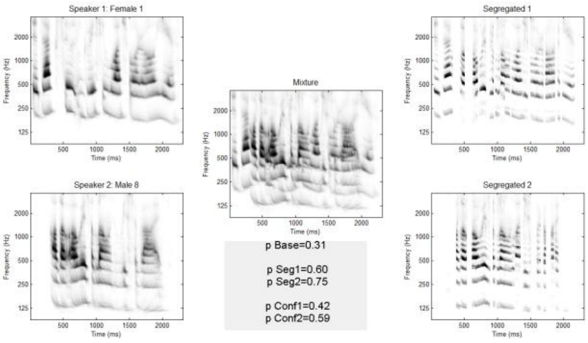

Figure 4. *An example of segregating a mixture of male and female speech signals*

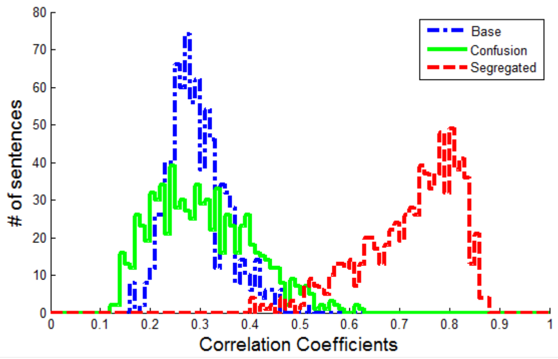

Figure 5. *The histogram of the correlation coefficients* ρ_{Base} (base), ρ_{Seg} (segregated) and ρ_{Conf} (confusion) *for 400 mixtures;* ρ_{Seg} is significantly higher than ρ_{Conf}.

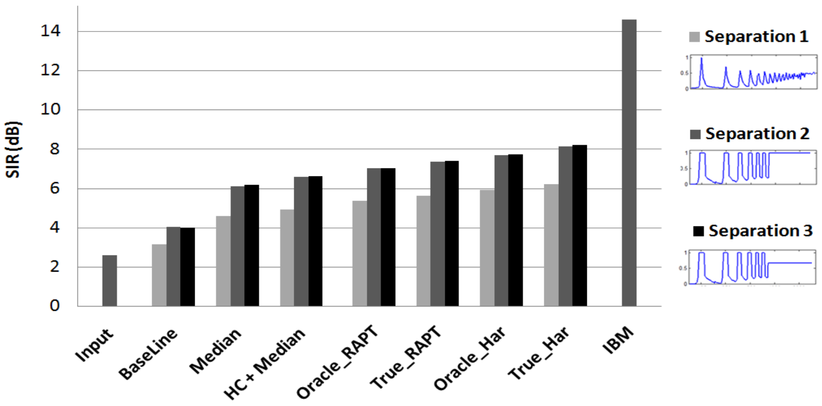

Figure 6. *SIR Results. Separation 1, 2 and 3 are performed by: original harmonic templates, six-first with fixed high frequency, and six-first with weighted high frequency sets, Figure2 (a), (b), and (c), respectively. Input and ideal binary mask (IBM) are the SIR boundaries. Proposed results are Median and HC+Median which show higher SIR than Baseline. Oracle and True results are also the maximum potential of the GCT and HMS in the proposed system, respectively.*

3.3. SIR Results

SIR_{ij} , the signal-to-interference ratio of the j^{th} output, O_j , when it estimates the i^{th} intput signal, S_i, is defined by

$$SIR_{ij} = 10log_{10}\frac{power\ (S_i)}{power\ (O_j/a_{ji} - S_i)} \qquad (8)$$

where a_{ji} is the scale factor and is calculated by

$$O_1 = a_{11}S_1 + a_{12}S_2 + \epsilon_1 \qquad (9)$$
$$O_2 = a_{21}S_1 + a_{22}S_2 + \epsilon_2.$$

To avoid the permutation for estimated S_1 and S_2, we used max(SIR_{11}, SIR_{12}) and max(SIR_{21}, SIR_{22}), respectively. The SIR is calculated in auditory spectrogram domain. The average SIR on 90 mixtures of male and female signals in TIMIT database is shown in Figure 6. The results belong to the dominant speaker in each mixture. Separation 1, 2 and 3 are performed by different harmonic templates sets, shown in Figure 2(a), (b) and (c), respectively. In Figure 6, SIR results of the input and the ideal binary mask (*IBM*) [8] are minimum and maximum boundaries, respectively. Baseline system shows an average of 1.5 dB improvement; while the proposed model shows 3.6 dB for *Median* and 4 dB for *HC+Median* selections.

More experiments are performed to show the potential of the GCT-based multi-pitch extraction and the maximum performance of our HMS separation. Therefore, it cannot be achieved in practice, but shows the upper limits of the proposed methods for multi-pitch extraction and HMS. The *Oracle* results come from the oracle pitch selection. The *oracle* pitch is defined as the closest value to the true pitch among the whole candidates. The *True* results come from the HMS with the true pitch values. To calculate the true pitch values from clean speech signals, two approaches are utilized in this paper; the harmonic-based approach presented in the baseline system, named *Har*, and the conventional *RAPT* approach [9]. In the both true pitch extractions, the difference of *Oracle* and *True* results are less than 0.4 dB. However, our proposed *Har* pitch extraction shows 0.8 dB higher SIR than the *RAPT*. Generally, the maximum average SIR is 8.2 dB, which is still far from the 14.6 dB in *IBM*. Also, in every case, the separation 3 shows higher performance than separation 2, and separation 1, in order.

4. Conclusion

This paper proposed a CASA model based on a 2D spectro-temporal processing, called GCT analysis, and HMS. In this model, three biologically inspired approaches were utilized for monaural speech separation: to generate auditory spectrogram, to analyze its spectro-temporal content, and to track the harmonic structure. A baseline system is designed for comparison which is similar to the proposed model except for the spectro-temporal analysis. The experimental result demonstrated significant improvement compare to the baseline which implicates the effect of the spectro-temporal analysis, mimicking the function of the auditory cortex. The *True* and *Oracle* results of our harmonic-based pitch extraction show better SIR than the conventional pitch extraction *RAPT*, which is an evidence to convince our harmonic analysis structure and the harmonic templates sets. Also, the small gap between *Oracle* and *True* results show the strength of the GCT-based multi-pitch extraction. However, the gap between *HC+Median* and the *Oracle* shows the multi-pitch selection methods can be further improved. Also, comparing our results with the ideal case shows the proposed is far from the ideal binary mask and how much it must be improved.

5. Acknowledgements

This research was performed for the Intelligent Robotics Development Program, one of the Frontier R&D Programs funded by the Ministry of Knowledge Economy (MKE F0005000-2010-32).

6. References

[1] Elhilalia, M. and Shamma, Sh., “A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation”, J. Acoust Soc. Am. 124(6):3751-71, 2008.

[2] Alain, C., Reinke, K., McDonald, K.L., Chau, W., Tam, F., Pacurar, A. and Graham, S., “Left thalamo-cortical network implicated in successful speech separation and identification”, NeuroImage, 26(2):592-9, 2005.

[3] Wang, T.T. and Quatieri, T.F., “2-D Processing Of Speech For Multi-Pitch Analysis”, in proc. Interspeech, 2827-2830, 2009.

[4] Shamma, Sh. and Klein, D., “The case of the missing pitch templates: how harmonic templates emerge in the early auditory system”. J. Acoust Soc. Am. 107(5 Pt 1):2631-44, 2000.

[5] Wang K., and Shamma Sh., “Self-normalization and noise-robustness in early auditory representations”, IEEE Trans. Speech Audio Proc. 2(3): 421‐435, 1994.

[6] Stark, M., Wohlmayr, M. and Pernkopf, F., “Source-Filter-Based Single-Channel Speech Separation Using Pitch Information”, IEEE Trans. Audio, Speech & Language Proc. 19(2): 242-255, 2011.

[7] Ma, N., Green, P., Barker, J. and Coy, A., “Exploiting correlogram structure for robust speech recognition with multiple speech sources”, Speech Communication, 49(12):874–891, 2007.

[8] Wang, D., “On ideal binary mask as the computational goal of auditory scene analysis,” in P. Divenyi [Ed], Speech Separation by Humans and Machines, 181–197, Kluwer Academic, Norwell, Mass, USA, 2005.

[9] Talkin, D., “A robust algorithm for pitch tracking (RAPT),” in Speech coding and synthesis (Elsevier, Ed.), 495–518, 1995.