SINGLE CHANNEL SPEECH AND BACKGROUND SEGREGATION THROUGH HARMONIC-TEMPORAL CLUSTERING

Jonathan Le Roux[1,2]*, Hirokazu Kameoka*[1 †]*, Nobutaka Ono*[1]*, Alain de Cheveigné*[2]*, Shigeki Sagayama*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[2]CNRS, Université Paris 5, and Ecole Normale Supérieure, Paris, France
{leroux,kameoka,onono,sagayama}@hil.t.u-tokyo.ac.jp, Alain.de.Cheveigne@ens.fr

ABSTRACT

The design of effective algorithms for single-channel analysis of complex and varied acoustical scenes is a very important and challenging problem. We present here the application of the recently introduced Harmonic-Temporal Clustering (HTC) framework to single channel speech enhancement, background retrieval and speaker separation. HTC processing relies on a precise parametric description of the voiced parts of speech derived from the power spectrum. We explain the positioning of the algorithm inside the Computational Acoustic Scene Analysis (CASA) area, describe the theoretical background of the method, show through preliminary experiments its basic feasibility, and discuss potential improvements.

1. INTRODUCTION

The design of an effective method for the analysis of complex and varied acoustical scenes is a very important and challenging problem. Many applications, such as automatic speech recognition (ASR) or speaker identification, would for example benefit from the ability of such a system to reduce acoustic interferences which often occur simultaneously with speech in real environments. Being able to locate and extract a portion of an acoustical scene, or on the contrary to cancel it, would also lead to very appealing applications such as instrument separation inside a multipitch track, background music recovery or voice activity detection (VAD).

Although there exist general methods for signal separation or enhancement in multisensor frameworks, based for example on independent component analysis or spatial filtering, single-channel solutions are necessary for many applications, such as in telecommunication, analysis of monaural CD recordings, automatic news search or background music determination in television programs, for example. Implementing the single-channel separation problem in computers has proven to be extremely challenging. However, human listeners are able to concentrate on listening to a specific target sound without difficulty even in the situation where many speakers are talking at the same time, and they are are able to do so even in monaural situations. This fact has persuaded many scientists that the human auditory system has a significant ability to actively recognize the external environment, in other words to perform an *auditory scene analysis* (ASA) and has been attracting interest since Bregman's book was published [1]. In [1], Bregman has shown through experiments the psychological evidences that the auditory system segregates the acoustic signal into spectrogram-like pieces, called *auditory elements*, which are grouped into *auditory streams* according to several grouping cues. Recent efforts are being directed towards the reproduction of this ability of the auditory system in computers, in a framework called "Computational Auditory Scene Analysis (CASA)" [2, 3, 4, 5]. The main focus of today's CASA research is to develop a source separation method based upon the grouping cues suggested by Bregman. More specifically, the main purpose is to extract useful features (for example, the fundamental frequency F_0) or to restore the target signal of interest by performing the segregation process and grouping process through a computational algorithm.

Many methods utilizing the grouping cues have been proposed (see [5] for a list of references), in most of which the grouping process is usually implemented in two steps. Instantaneous features are first extracted at each discrete time point, which corresponds to the grouping process in the frequency direction, and a post-processing is then performed on these features to reduce errors and/or obtain continuous tracks, through hidden Markov model (HMM), multiple agents, or some dynamical system such as Kalman filtering, corresponding to the grouping process in the time direction. Considering that from an engineering point of view, it should be more efficient to perform the analysis in both time and frequency directions simultaneously, we formulated a unified estimation framework "Harmonic-Temporal Clustering" (HTC) for the two dimensional structure of time-frequency power spectra [5], in contrast to the conventional strategy.

In our previous works [5, 6], we presented an F_0 estimation algorithm based on HTC for both music and speech. We present here the next step in the application of HTC, single-channel speech signal processing for speech enhancement, background retrieval, and speaker separation. We use the estimated HTC models to build masking functions which we apply to the power spectra to extract speech, retrieve the background and separate two speakers.

2. INTRODUCTION OF HTC

We briefly introduce here the HTC framework. More details on the theory, the implementation and the performance of F_0 estimation through HTC can be found in [5, 6, 7].

2.1. General HTC method

Consider the wavelet power spectrum $W(x,t)$, where x is log-frequency and t is time, of a signal recorded from an acoustical scene. The problem is to approximate it as well as possible as the sum of K parametric source models $q_k(x,t;\mathbf{\Theta})$, where $\mathbf{\Theta}$ is the set of model parameters, modeling the power spectrum of K "objects" each with its own F_0 contour $\mu_k(t)$.

[†]Now with NTT Communication Science Laboratories, NTT Corporation, Atsugi, Japan

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

As described in [5], each source model is expressed as a Gaussian Mixture Model with constraints on the characteristics of the kernel distributions: supposing there is harmonicity with N partials modeled in the frequency direction, and that the power envelope is described using Y kernel functions in the time direction, we can rewrite each source model as

$$q_k(x,t;\mathbf{\Theta}) = \sum_{n=1}^{N}\sum_{y=0}^{Y-1} S_{kny}(x,t;\mathbf{\Theta}), \quad (1)$$

with kernel densities $S_{kny}(x,t;\mathbf{\Theta})$ which are assumed to have the following shape:

$$S_{kny}(x,t;\mathbf{\Theta}) \triangleq \frac{w_k v_{kn} u_{kny}}{2\pi\sigma_k\phi_k} e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}}, \quad (2)$$

where the parameters w_k, v_{kn} and u_{kny} are normalized to unity. A graphical representation of a HTC source model $q_k(x,t;\mathbf{\Theta})$ can be seen in Fig. 1.

2.2. Speech modeling

In order to model the spectrum of a speech utterance, we make several assumptions. First, we suppose that the F_0 contour is smooth and defined on the whole interval: we will not make voiced/unvoiced decisions, and F_0 values are assumed continuous. Previously, HTC was applied to musical signals with piecewise constant F_0 contours. For speech, it is necessary to model continuously varying contours. For that, we chose to use cubic spline functions as a general class of smooth functions. To model the variation over time of the shape of the spectral envelope, the speech segment is modeled as a succession in time of slightly overlapping source models sharing a common F_0 contour. In the single speaker case, all the source models inside the HTC model share the same F_0 contour: $\mu_k(t) = \mu(t), \forall k$, while for multiple speakers, according to the number of F_0 contours that we want to estimate, we group source models into subsets sharing a common F_0 contour. As the structure is assumed harmonic, the model takes advantage of the voiced parts of the speech utterance.

It is shown in [5] and [6] that prior distributions can be introduced on the parameters, and that the HTC model can be efficiently optimized using an EM-like algorithm to minimize the "distance" between the parametric HTC model and the observed spectrogram measured by the $\mathcal{I}$-divergence between them.

2.3. Noise Modeling

We introduce a noise model to cope with the broadband background noise which can be a disturbance in the process of clustering the harmonic portions of speech. The idea to design this model was that detecting the harmonic parts of the spectrogram in a noisy background corresponds to searching for thin and harmonically distributed "islands" which emerge from a "sea" of noise. We thus chose to model the noise using a mixture of Gaussian distributions with large fixed variance and with centers fixed on a grid, the only parameters being the mixing weights and the ratio of noise power inside the whole spectrogram. The noise model parameters are optimized simultaneously with the other parameters in the same EM-like framework [6].

3. SPECTROGRAM MODIFICATION

Speech enhancement, background retrieval and speaker separation can be performed very simply through HTC, based on the classic idea of masking function, often used in CASA oriented methods.

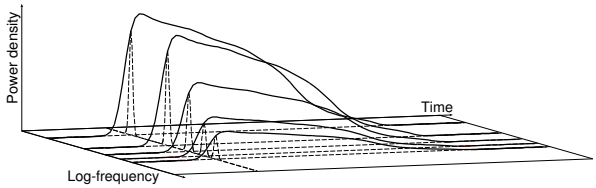

Figure 1: Profile of a HTC source model $q_k(x,t;\mathbf{\Theta})$

3.1. Ratio of HTC models

When the noise to be cancelled is broadband, an estimation of the spectrogram where the interference has been cancelled can be obtained from the original spectrogram by looking, at each point (x,t), at the proportion of the "clean" part inside the whole parametric model. In the same way, in the multiple speakers case, when several speech models are used simultaneously, the ratio of one speech model inside the whole at each time-frequency bin can be used as a mask. This corresponds to the E-step in the EM algorithm. In the course of the optimization of the model, it is actually on this "cleaned" part of the spectrogram that the F_0 contour estimation is performed, during the M-step, enabling the F_0 estimation to perform well even in very noisy environments or within multiple speakers [6].

3.2. Direct use of HTC models

However, when the noise to be cancelled is not a broadband noise, we can expect that the noise model will not be able to cope totally with the background noise. It might thus as well happen that the total power of the noise model is very low in absence of broadband noise, and using the ratio of speech model in the whole would then make less sense, as this ratio would almost always be close to 1, thus leading to a very ineffective mask. Introducing noise models to cope with more types of background noise could be a way to deal with that problem, but it is limited by several problems: background noises could be of infinitely many kinds, and the multiplication of models would lead to a larger computation cost, and might also conflict with the estimation of the speech model. It might thus be simpler and more effective to look directly at the estimated speech model itself to build a masking function. The speech model has been designed to encompass the acoustic characteristics of speech, and has by construction a harmonic structure.

To use the speech model as a mask, we use a "filtered" version of the speech model which broadens its peaks:

$$\tilde{Q}(x,t) = \frac{1}{1+(\frac{\epsilon}{Q(x,t)})^p}, \quad (3)$$

where Q is the speech model, normalized such that its maximum is 1, ϵ a small constant, typically between 10^{-3} and 10^{-1}, and p a constant which tunes the broadening of the peaks.

The interference can also be retrieved using a masking function obtained through HTC. We simply consider $1-\tilde{Q}$, where $\tilde{Q}$ is defined as in (3), as a mask function to apply to the noisy spectrogram to retrieve the interference part from the mixture.

In all cases, the modified power spectrogram is coupled with the phase of the noisy spectrogram to obtain an estimation of the denoised complex spectrogram. An inverse transform is then used to synthesize the denoised signal back.

Table 1: *SNR results (dB) for the enhanced speech.*

	n0	n1	n2	n3	n4	n5	n6	n7	n8	n9
Mixture	-3.27	-4.08	10.20	4.39	4.05	-5.83	1.90	6.57	10.53	0.75
Hu-Wang	16.34	7.83	16.71	8.32	10.88	14.41	16.89	11.97	14.44	5.27
$p = 1, \epsilon = 0.005$	-3.98	-0.54	15.11	6.63	4.90	-5.76	1.91	7.01	10.65	0.69
$p = 1, \epsilon = 0.1$	-6.04	5.61	11.39	7.79	6.69	-5.31	4.20	7.66	9.86	-0.47
$p = 2, \epsilon = 0.1$	-6.51	7.61	9.56	7.24	6.53	-5.57	4.72	7.06	8.68	-0.98

3.3. STFT and Wavelet Spectrograms

The HTC models are optimized to fit the wavelet power spectrum of an utterance. The basic idea to synthesize an enhanced or denoised speech, is, as described above, to modify the wavelet power spectrum and use an inverse wavelet transform. So far, a Gabor transform has been used for HTC analysis in [5] and [6], and a first option is to perform the analysis-synthesis with this transform. As argued in [7], the HTC framework is naturally designed to fit a power spectrogram obtained with a constant-Q filterbank based on an analyzing wavelet whose Fourier transform is of the form

$$\Psi(\omega) = \Psi^*(\omega) = \begin{cases} \exp\left(- \frac{\left(\log \omega\right)^2}{4\sigma^2} \right) & (\omega > 0) \\ 0 & (\omega \leqq 0) \end{cases}. \quad (4)$$

One can thus also try to perform the analysis-synthesis with this analysing wavelet.

But as the models obtained on the wavelet spectrogram have parametric expressions, it is also possible to generate masking functions in the linear-frequency domain, and simply use STFT and inverse STFT, for example by overlap-add method, to generate a modified speech or background. Although this process is expected to lead lower-quality results than wavelet-based synthesis due to misfitting between the analysis and synthesis methods, it is much faster, and still gives interesting preliminary results on the basic performance of the method.

3.4. Relation to adaptive comb filtering methods

Comb filtering first requires a robust F_0 estimation, and according to [4], suffers from the fact that it retains too much interference as it passes through all frequency components close to the multiples of target F_0. The HTC frameworks, on the opposite, features an embedded estimation of the F_0, and estimates separately the powers and shapes of the harmonics. HTC also estimates the parameters simultaneously in the time and frequency directions, thus integrating continuity in the model and making it more robust, on the contrary to comb filtering methods. In the multiple speaker case especially, if the speech of two speakers are harmonically related, a comb filter method will inevitably fail, while HTC can still perform well, as shown in Section 4.3.

4. EXPERIMENTS

We performed three types of experiments to confirm the basic effectiveness of our method for speech enhancement, background retrieval and speaker separation, using STFT spectrograms. We used the SNR as a quantitative measure of the performance of our algorithm, and tested different settings for the mask functions. We shall stress the fact that the SNR, although it is an easy-to-compute objective value, may not give a full idea of the performance of a CASA system, and may especially differ significantly from a perceptive evaluation. The human ear tends to prefer a stronger masking, even if it introduces artifacts or slightly modifies speech, which is not advantageous in an SNR way.

4.1. Speech enhancement

We used a corpus of 100 mixtures of voiced speech and interference [2], commonly used in CASA research. There are 10 interferences: n0, 1-kHz pure tone; n1, white noise; n2, noise bursts; n3, "cocktail party" noise; n4, rock music; n5, siren; n6, trill telephone; n7, female speech; n8, male speech; and n9, female speech. The results are shown in Table 1, for different mask settings on various types of interferences. Each value in the table represents the average SNR for one interference mixed with 10 target utterances. Mixture designates the Signal-to-Interference Ratio in the original mixture, and Hu-Wang stands for the state-of-the-art algorithm presented in [4]. The HTC enhanced speech is generated using Eq. (3) with p and ϵ as indicated in the first column. 40 iterations of HTC were performed, and the parameters were as in [6], with a speech model and a noise model, apart from the number of harmonics considered, which was set to 40. We note that according to the type of interference, different settings lead to better results. This is due to the fact that these settings change the sharpness of the peaks of the mask, introducing a trade-off between Signal-to-Interference Ratio (SIR) and Signal-to-Artifact Ratio (SAR). According to the acoustical properties of the interference to reduce, it is thus possible to use different values for the parameters.

For n1, n2, n3 and n4, the results are promising, with results close to the Hu-Wang algorithm for the first three. It is so far less effective on the other interferences. For n0 and n5, which are interferences with a strong localized signal overlapping with harmonics of the speech utterance, our method failed as the speech model mistook the interference for a harmonic and rose the power of the corresponding Gaussian functions. This should be dealt with in the future, for example by using a stronger constraint on the power of the harmonics and on the shape of the power envelope in the time direction. The analysis of SIR and SAR results tends to show that our algorithm reduces the interference very effectively, but creates artifacts. The question whether these artifacts are perceptually significant or not should be further investigated, as is the improvement of the quality of the synthesized speech, especially using inverse wavelet transforms. We shall note in particular that, contrary to the algorithm by Hu and Wang, our method does not seem to generate musical noise.

4.2. Speech cancellation for background enhancement

The HTC framework can be used not only for speech enhancement, but also for speech cancelling and background retrieval, as explained in section 3.2. There are very interesting potential applications to this task, such as the retrieval of speech in the background, or background music retrieval. This last issue is of particular importance: in an acoustical scene where someone is speaking with music playing in the background, being able to "clean" the background music from the speech would ease the automatic

Table 2: *SNR results (dB) for the retrieved background.*

	n3	n4	n7	n8	n9
Mixture	-4.39	-4.05	-6.57	-10.53	-0.75
$p = 3, \epsilon^3 = 5.10^{-4}$	0.90	-0.01	-2.99	-4.85	-6.44
$p = 3, \epsilon^3 = 2.10^{-3}$	0.39	0.28	-2.62	-5.12	-5.36

recognition of copyrighted material inside television programs for example. Another interesting application is the automatic cancellation of the vocal part inside a music piece to produce karaoke accompaniments at lower cost and from the real song, thus with a better quality than the usual MIDI accompaniments.

The results for the interferences which retrieval is of potential interest for applications are presented in Table 2. Mixture designates the Interference-to-Signal Ratio in the original mixture. To our knowledge, no results by previous methods are available on this task. Background is retrieved through Eq. (3) with p and ϵ as indicated. A substantial increase in SNR could be obtained for n3, n4, n7 and n8. For an interference constituting of other speech with close average power such as n9, we will see in the next section that using two speech models leads to better results.

4.3. Speaker Separation

By using two speech models, we showed in [6] that the F_0 contours of concurrent speech by two speakers with close average power can be effectively estimated through HTC. As explained in Section 3.1, the speech of each speaker can be reconstructed using the proportion of its model in the whole. We performed a separation experiment on two mixtures. On the first one, v0n9 from Cooke's database, we obtained an improvement from an initial SNR of 0dB to 6dB for both speakers. This is a very difficult task as the harmonics of the speakers almost constantly overlap. The clean spectrograms of the utterances v0 (male speaker) and n9 (female speaker) can be seen in Fig. 2 and Fig. 3 respectively, and their mixture in Fig. 4. The corresponding spectrograms extracted using HTC from the mixture v0n9 can be seen in Fig. 5 and Fig. 6 respectively. An exponent 0.3 was used as non-linear scaling for the figures. The second mixture constitutes of a male Japanese speaker uttering "aoi" and a female Japanese speaker uttering "oi wo ou", the female speaker being 5dB stronger. We obtained an improvement from the initial -5dB to $+3$dB for the male speaker, and from the initial $+5$dB to 9.3dB for the female speaker.

5. CONCLUSIONS

We presented several applications of the HTC framework for single channel speech signal processing problems and showed its basic effectiveness for tasks as various as speech enhancement, speech cancellation for background retrieval, with potential applications in background music retrieval, and speaker separation. Future works include the improvement of the quality of the output sound by using wavelet transforms instead of STFT ones, a thorough study of the performance of HTC compared to previous works, both on voiced speech and on data containing unvoiced speech, the use of auditory-based distortion measures and an evaluation through subjective listening tests.

6. REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, Cambridge, 1990.

[2] M. P. Cooke, "Modeling Auditory Processing and Organisation," Ph.D. dissertation, University of Sheffield, 1993.

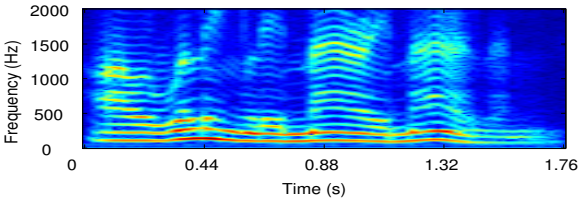

Figure 2: *Clean spectrogram of speaker v*0

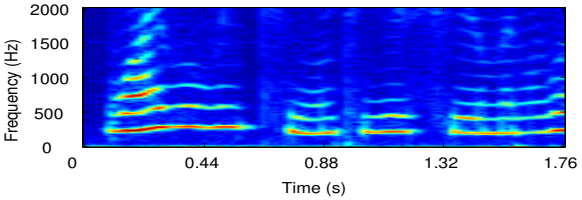

Figure 3: *Clean spectrogram of speaker n*9

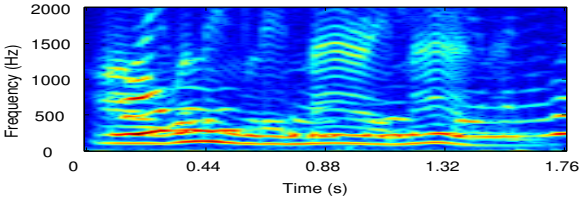

Figure 4: *Spectrogram of the mixture v*0*n*9

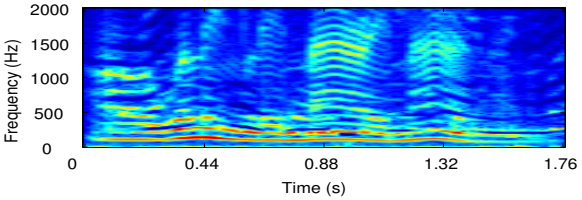

Figure 5: *Estimated spectrogram of speaker v*0

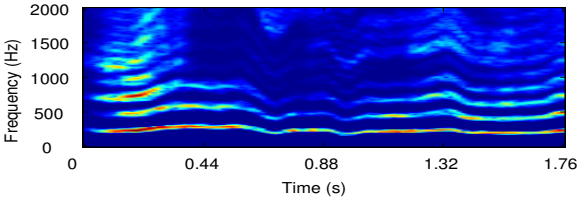

Figure 6: *Estimated spectrogram of speaker n*9

[3] D. Ellis, "Prediction-driven Computational Auditory Scene Analysis," Ph.D. dissertation, MIT, 1996.

[4] G. Hu and D. Wang, "Monaural speech segregation. based on pitch tracking and amplitude modulation," *IEEE. Trans. Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[5] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," in *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, 2007, pp. 982–994.

[6] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Single and multiple F_0 contour estimation through parametric spectrogram modeling of speech in noisy environments," in *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, 2007, pp. 1135–1145.

[7] H. Kameoka, "Statistical Approach to Multipitch Analysis," Ph.D. dissertation, The University of Tokyo, 2007.