Monaural speech separation through Harmonic-Temporal Clustering of the power spectrum *

◎ Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono (The University of Tokyo), Alain de Cheveigné (CNRS, ENS, Paris 5) and Shigeki Sagayama (The University of Tokyo)

1 Introduction

We present here the application of the recently introduced Harmonic-Temporal Clustering (HTC) framework to single channel speech separation. HTC processing relies on a precise parametric description of the voiced parts of speech derived from the power spectrum, the F_0 contour being modeled as a spline contour. This framework has been introduced in [1], where more details can be found, with application to robust F_0 estimation. We show here that an analytical update equation can be derived for the spline contour parameters. We also present results of a simple evaluation experiment which shows the effectivity of our method on the separation of concurrent speech by two speakers with close average power.

2 Spline parameter update equation

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 speech models $q_k(x,t;\mathbf{\Theta})$, where $\mathbf{\Theta}$ is the set of model parameters, modeling the power spectrum of K speakers each with its own F_0 contour $\mu_k(t)$. We use cubic spline functions as a general class of smooth functions to model the F_0 contour. The F_0 contours of all speakers can be updated simultaneously, but independently from each other, and one can thus assume without loss of generality here that there is only one speaker.

2.1 Spline contour

The analysis interval is divided into subintervals $[t_i, t_{i+1})$ of equal length ϵ. The parameters of the spline contour model are then the values z_i of the F_0 at each bounding point t_i. Assuming that the second derivative vanishes at the bounds of the analysis interval leads to the so-called natural splines. An analytical expression for the contour $\mu(t;\boldsymbol{z})$ as a concatenation of third order polynomials can be classically obtained, which can be noticed to be linear in $\boldsymbol{z}$:

$$\mu(t;\boldsymbol{z}) = \boldsymbol{A}(t)^T \boldsymbol{z} \tag{1}$$

where $\boldsymbol{A}(t)$ is a column vector which elements are, for $t \in [t_i, t_{i+1})$, third order polynomials in t. We note furthermore that

$$\boldsymbol{A}(t) = \nabla_{\boldsymbol{z}}\, \mu(t;\boldsymbol{z}). \tag{2}$$

2.2 Optimization of the objective function

During the M-step of the EM algorithm, one wants to maximize with respect to $\mathbf{\Theta}$ the quantity

$$\mathcal{J}(\mathbf{\Theta}) \triangleq \iint_D \Big(\sum_{k,n,y} \ell_{kny}(x,t) \log \frac{S_{kny}(x,t;\mathbf{\Theta})}{m_{kny}(x,t)} \Big) \mathrm{d}x\mathrm{d}t,$$

where

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

is the parametric expression of a single kernel density, $Q = \sum_{k,n,y} S_{kny}$ is the modeled spectrogram, $W(x,t)$ the observed spectrogram, $\ell_{kny}(x,t) = m_{kny}(x,t)W(x,t)$ with $m_{kny}(x,t)$ the membership degrees obtained during the E-step (for more details, please refer to [1]).

We compute the gradient with respect to $\boldsymbol{z}$:

$$\nabla_{\boldsymbol{z}}\mathcal{J} = \iint_D \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2} (x - \boldsymbol{A}(t)^T\boldsymbol{z} - \log n)\boldsymbol{A}(t)\mathrm{d}x\mathrm{d}t.$$

Let

$$\begin{aligned} \phi(t) &= \int_{D_x} \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2} (x - \log n)\mathrm{d}x, \\ \gamma(t) &= \int_{D_x} \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2} \mathrm{d}x. \end{aligned}$$

Then

$$\nabla_{\boldsymbol{z}}\mathcal{J} = \int \phi(t)\boldsymbol{A}(t)\mathrm{d}t - \left(\int \gamma(t)\boldsymbol{A}(t)\boldsymbol{A}(t)^T \mathrm{d}t \right) \boldsymbol{z}$$

Putting to 0 the gradient w.r.t. $\boldsymbol{z}$, one can find the update equation for $\boldsymbol{z}$:

$$\boldsymbol{z} = (H_{\boldsymbol{z}}\mathcal{J})^{-1} \int \phi(t)\boldsymbol{A}(t)\mathrm{d}t, \tag{3}$$

where $H_{\boldsymbol{z}}\mathcal{J} = -\int \gamma(t)\boldsymbol{A}(t)\boldsymbol{A}(t)^T\mathrm{d}t$ is the Hessian matrix, which is at least negative semi-definite, thus ensuring that we are indeed looking at a maximum of the objective function.

3 Separation of co-channel concurrent speech

By using two speech models, we showed in [1] that the F_0 contours of concurrent speech by two speakers with close average power can be effectively estimated through HTC. When several

* パワースペクトルの調波時間構造化クラスタリングによるモノラル音声分離、ルルー・ジョナトン、亀岡弘和、小野順貴 (東大情報理工)、ドゥシュベニエ・アラン (CNRS/ENS/Paris 5)、嵯峨山茂樹 (東大情報理工)

Table 1 *SNR results (dB) for the speaker separation.*

	v0n9	v1n9	v2n9	v3n9	v4n9	v5n9	v6n9	v7n9	v8n9	v9n9	Average
Mixture SNR	0.30	-1.98	-0.96	2.92	0.63	0.80	-0.05	2.12	1.38	2.72	0.79
Hu-Wang	3.89	2.84	7.59	8.12	3.51	3.24	4.00	6.97	4.84	8.69	5.37
HTC, Speaker 1	7.26	7.40	8.67	10.61	6.64	9.80	7.01	9.58	7.14	12.32	8.64
Mixture RNS	-0.30	1.98	0.96	-2.92	-0.63	-0.80	0.05	-2.12	-1.38	-2.72	-0.79
HTC, Speaker 2	6.33	9.00	9.03	6.82	5.13	8.83	6.20	7.04	5.03	8.49	7.19

speech models q_k are used simultaneously, the ratio $q_i(x,t;\mathbf{\Theta}_{opt})/\sum_k q_k(x,t;\mathbf{\Theta}_{opt})$ of one speech model inside the whole at each time-frequency bin can be used as a mask to recosntruct the speech of each speaker. We performed a separation experiment on ten mixtures from Cooke's database (see [1] for details on the database and HTC experimental setup), where utterances by male speakers v0 to v9 are mixed with an utterance by a female speaker n9. The SNR is close to 0dB for each utterance, and we note that this is a very difficult task as for some of the mixtures 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. 1 and Fig. 2 respectively, and their mixture in Fig. 3. The corresponding spectrograms extracted using HTC from the mixture v0n9 can be seen in Fig. 4 and Fig. 5 respectively. A constant-Q filterbank transform and inverse transform was used.

We used the SNR as a quantitative measure of the performance of our algorithm. The results for the ten mixtures are shown in Table. 1. For comparison, SNR results obtained using Hu and Wang's state-of-the-art algorithm [2] are also given. This algorithm only focuses on the target source, and results for the second speaker are thus not available. One can see that our method outperforms this algorithm on this task. We also note that, contrary to the algorithm of Hu and Wang, it does not seem to generate musical noise.

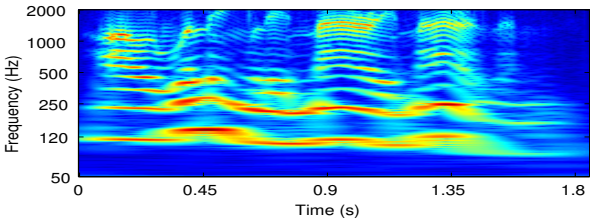

Fig. 1 *Clean spectrogram of speaker v0*

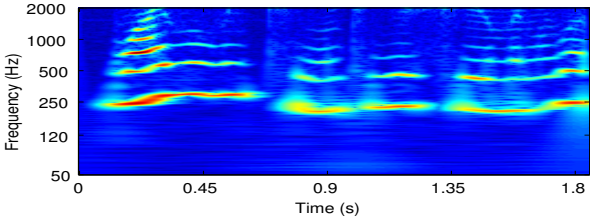

Fig. 2 *Clean spectrogram of speaker n9*

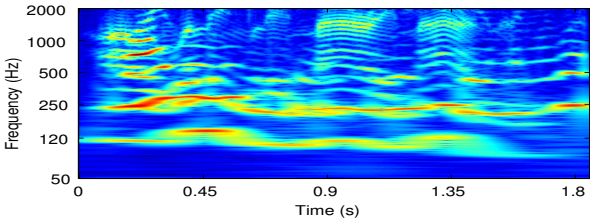

Fig. 3 *Spectrogram of the mixture v0n9*

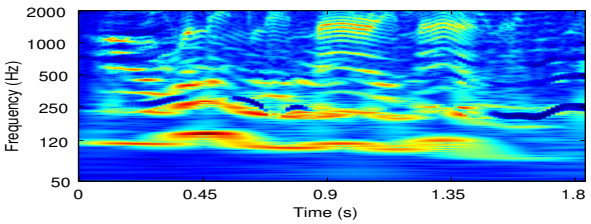

Fig. 4 *Estimated spectrogram of speaker v0*

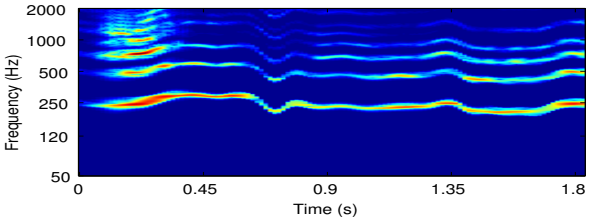

Fig. 5 *Estimated spectrogram of speaker n9*

4 conclusion

We presented a new algorithm for single channel speech separation based on HTC, and showed through experimental evaluation that it outperforms previous work on the separation of concurrent voiced speech by male and female speakers with close average power. We plan to perform experiments on normal speech as well as for two male or two female speakers.

References

[1] 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 Proc.*, 2007, vol. 15, pp. 1135–1145.

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