Harmonic Temporal Clustering of Speech Spectrum *

◎ Jonathan Le Roux, Hirokazu Kameoka,
Nobutaka Ono and Shigeki Sagayama (The University of Tokyo)

1 Introduction

We introduce in this paper the extension to speech spectrum of the Harmonic-Temporal Structured Clustering (HTC) method [1] developed for feature extraction of multi-stream music signal and estimating the structure in time and frequency directions simultaneously by decomposing the energy pattern into distinct clusters such that each of them is originated from a single sound stream. We first consider here the case of a single speaker, and impose a common pitch contour function to all the clusters which are now intended at representing portions with different spectral structures succeeding to each other continuously. By using a spline model for the pitch, we are able to obtain analytical update equations and optimize the model using the EM algorithm. As a first insight on this model's accuracy, we present very good results on its performance as a pitch extractor.

2 Formulation of the model

2.1 Original HTC model

The problem is to approximate as well as possible an observed power spectrum $W(x,t)$, where x and t are log-frequency and time, by the sum of K parametric models $q_k(x,t;\mathbf{\Theta})$ modeling the power spectrum of K "sources", with pitch $\mu_k(t)$, written in the form

$$q_k(x,t;\mathbf{\Theta}) = \sum_{n,y} S_{kny}(x,t;\mathbf{\Theta}), \tag{1}$$

where $\mathbf{\Theta}$ is the set of all parameters and with logarithmic kernel densities $\log S_{kny}(x,t;\mathbf{\Theta})$ which are supposed to have the following shape:

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

where the parameters w_k, v_{kn} and u_{kny} are normalized to unity (Refer to [1] for a precise description of the parameters). The S_{kny} are subkernels of energy for which a membership degree $m(k,n,y;x,t)$ among the total power spectrum W is introduced, such that the energy for each subkernel is

$$l_{kny}(x,t) = m(k,n,y;x,t)W(x,t). \tag{3}$$

Our goal is then to minimize with respect to $\mathbf{\Theta}$ the sum over k,n and y of the inter-class Kullback-Leibler distance between $m(k,n,y;x,t)W(x,t)$ and $S_{kny}(x,t;\mathbf{\Theta})$. One can show [1] that this amounts to finding $\widehat{\mathbf{\Theta}}$ maximizing

$$\mathcal{I}(\mathbf{\Theta}) = \sum_{k,n,y} \iint_D l_{kny}(x,t)\log S_{kny}(x,t;\mathbf{\Theta})\mathrm{d}x\mathrm{d}t. \tag{4}$$

This fuzzy clustering problem can be solved using the EM algorithm, the possibility to obtain analytical update equations during the M-step depending on the actual expression of $\mu_k(t)$. If the HTC parameters do not enter in this expression, then the update equations obtained in [1] can be used as is, and we only need to obtain update equations for the pitch contour parameters.

2.2 HTC model for a single speaker

We suppose in this paper that there is only one speaker, and that the pitch is thus the same for all the sources inside the HTC model ($\mu_k = \mu$). Our intention is to have a succession in time of slightly overlapping sources which correspond if possible to successive phonemes, or at least to the main structures of the speech flow. As the structure is assumed harmonic, the model is originally designed for voiced speech.

We chose to use cubic splines to model a smooth speech pitch contour in order to obtain analytical update equations for the pitch parameters. The analysis interval is divided into subintervals $[t_i, t_{i+1}[$ of equal length and areas without significant sound where avoided using a threshold on the total energy. The parameters of the spline contour model are then the values $z_i = \mu(t_i)$ of the contour at t_i, and the values $z_i'' = \mu''(t_i)$ of the second derivative are obtained through $\boldsymbol{z}'' = M\boldsymbol{z}$ for a certain matrix M computed offline, supposing that the first-order derivative is 0 at the bounds of the analysis interval. The contour $\mu(t;\boldsymbol{z})$ on the whole interval is then given, for $t \in [t_i, t_{i+1}[$, by

$$\mu(t;\boldsymbol{z}) = \frac{1}{t_{i+1}-t_i}\Big(z_i(t_{i+1}-t) + z_{i+1}(t-t_i) - \frac{t-t_i}{6}(t_{i+1}-t)\big[(t_{i+2}-t)z_i'' + (t-t_{i-1})z_{i+1}''\big]\Big). \tag{5}$$

Pluging this expression into (2) and putting the derivatives with respect to the z_j to 0, one finds update equations analytically, as in [1]:

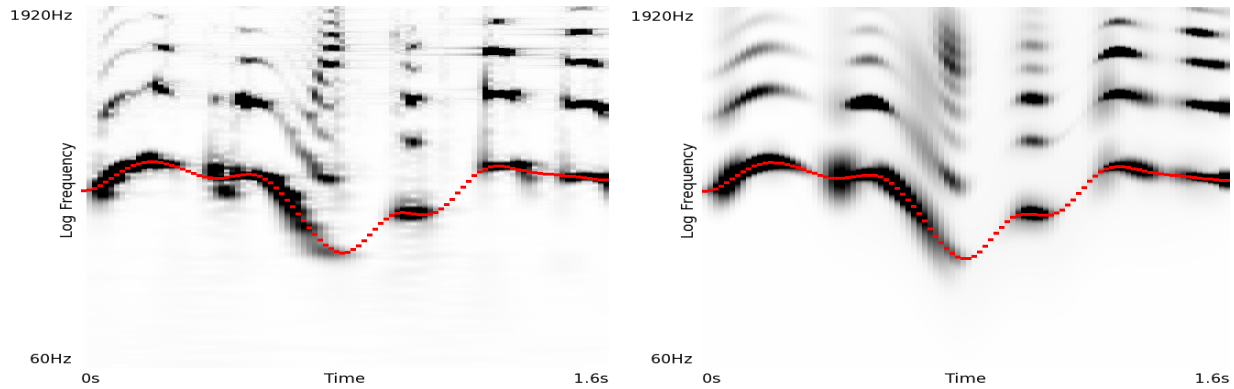
$$z_j^{n+1} = \frac{\sum_{k,n,y}\iint_D \Big(x-\hat{\mu}_j(t;\boldsymbol{z}^n)-\log n\Big) l_{kny}(x,t)\mathrm{d}x\mathrm{d}t}{\sum_{k,n,y}\iint_D \left(\frac{\partial\mu}{\partial z_j}(t;\boldsymbol{z}^n)\right)^2 l_{kny}(x,t)\mathrm{d}x\mathrm{d}t} \tag{6}$$

where $\hat{\mu}_j(t;\boldsymbol{z}^n) = \mu(t;\boldsymbol{z}^n) - \frac{\partial\mu}{\partial z_j}(t) z_j^n$ is the part of the pitch contour that does not depend on z_j.

The partial derivatives with respect to the other parameters (w_k, τ_k, u_{kny}, v_{kn}, ϕ_k, σ_k) are the same as in [1], as mentioned above.

An example is presented in Figure 1, on the Japanese sentence "Tsuuyaku denwa kokusai kaigi jimukyoku desu" ("通訳電話国際会議事務局です") uttered by a female speaker, with the observed and modeled (after 30 iterations) spectrogram. One can see that the model approximates very well the spectrum and that pitch contour is accurately estimated.

* 調波時間構造化クラスタリングによる音声分析、ルルー・ジョナトン、亀岡弘和、小野順貴、嵯峨山茂樹 (東大情報理工)



(a) Observed Spectrogram (b) Modeled Spectrogram

Fig. 1 Observed and modeled spectrograms with estimated pitch contour (「通訳電話国際会議事務局です」- "Tsuuyaku denwa kokusai kaigi jimukyoku desu", female speaker).

3 Experimental Evaluation

We evaluated the accuracy of the pitch estimation of our model on speech data from the B set of the ATR speech database [2], with the pitch labels given as a reference. We used 5 utterances of the male speaker MYI and 5 utterances of the female speaker FYM. Power spectrum was built using Gabor wavelet transform (16ms time resolution, 16kHz sampling rate, 60kHz lowest frequency, 14 cent frequency resolution). The spline contour's initial shape was fixed flat, at 280Hz for female speech and 230Hz for male speech, and the length of the interpolation intervals was fixed to 8 frames. The HTC model was built using $K = 8$ to 16 sources depending on the length of the utterance, each of them with $N = 10$ harmonics (although insufficient for proper analysis of a speech signal, this is enough for pitch extraction), and with power envelope functions made using $Y = 5$ gaussian kernels. The initial values of w_k, τ_k and ϕ_k were determined uniformly, and σ_k was fixed to 360 cent.

Deviations over 5% from the references were deemed to be gross errors and the areas were reference pitch is zero (no sound or unvoiced portions) were not considered in the computation of the accuracy. The results can be seen in Table 1, with for comparison the results we obtained using the classical cepstrum technique [3]. Although our model is not dedicated to pitch estimation, its accuracy is very good. Moreover, its harmonic nature doesn't hinder it from performing well even in the presence of unvoiced portions of speech inbetween the voiced portions where pitch is extracted, confirming the robustness of the 2D time-frequency analysis.

Table 1 *Pitch estimation results*

Speech File	Accuracy (%)	
	Cepstrum	Proposed
'myisda01'	88.2	95.4
'myisda02'	88.4	98.7
'myisda03'	84.8	97.9
'myisda04'	85.1	94.4
'myisda05'	76.8	98.1
'fymsda01'	86.3	96.8
'fymsda02'	87.1	92.8
'fymsda03'	83.3	96.6
'fymsda04'	86.7	97.5
'fymsda05'	85.2	98.7

4 Conclusion and future works

We introduced a model describing the spectrum as a sequence of spectral cluster models governed by a common pitch contour function, with smooth transitions in the temporal succession of the spectral structures. We explained how to optimize its parameters efficiently and evaluated the accuracy of the pitch contour estimation with very good results.

So far, the model has been designed for a single speaker, but we plan to extend it to simultaneous speech from multiple speakers, in order to separate them or extract their respective speech attributes, as well as multi-pitch analysis of music signals. Concerning the applications of our model, we would like to use the speech attributes obtained together with the ASAT framework [4] in the near future, and hope that these new features will help raise the recognition accuracy. From a specifically speech oriented point of view, we are currently working on introducing the Fujisaki pitch generation model [5] into our framework, as this should enable us to obtain even more relevant features to be used in the future.

References

[1] H. Kameoka, T. Nishimoto, and S. Sagayama. Harmonic-temporal-structured clustering via deterministic annealing em algorithm for audio feature extraction. In *Proc. ISMIR*, 2005.

[2] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis. *Speech Communication*, 9:357–363, 1990.

[3] A.-M. Noll. Cepstrum pitch determination. *J. Acoust. Soc. Amer.*, 41(2):293–309, February 1967.

[4] C.-H. Lee. From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition. In *Proc. ICSLP*, 2004.

[5] H. Fujisaki and S. Nagashima. A model for synthesis of pitch contours of connected speech. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, 28:53–60, 1969.