HARMONIC-TEMPORAL CLUSTERING OF SPEECH FOR SINGLE AND MULTIPLE F_0 CONTOUR ESTIMATION IN NOISY ENVIRONMENTS

Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné † and Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{leroux,kameoka,onono,sagayama}@hil.t.u-tokyo.ac.jp

†CNRS, Université Paris 5, Ecole Normale Supérieure,
Alain.de.Cheveigne@ens.fr

ABSTRACT

We present in this paper a novel F_0 contour estimation method based on a parametric description of the wavelet power spectrum of speech that accounts for its structure simultaneously in time and frequency directions. We model the speech spectrum as a sequence of spectral clusters governed by a smooth common F_0 contour expressed as a spline curve. The harmonic and temporal structure of these clusters and their common F_0 contour are estimated simultaneously. Through experimental comparisons with existing methods, we show that our algorithm is competitive on clean single-speaker speech, and that it outperforms existing methods both in the presence of noise and for the estimation of multiple F_0 contours of co-channel concurrent speech.

Index Terms— acoustic scene analysis, multi-pitch estimation, harmonic-temporal structured clustering (HTC), noisy speech, spline F_0 contour

1. INTRODUCTION

The design of an algorithm for the robust estimation of the F_0 contour of harmonic signals such as speech is a challenging problem which has been widely investigated [2, 3] but not yet solved satisfactorily. An algorithm that would perform with high accuracy in a wide range of background noises (white noise, pink noise, noise bursts, music, other speech...), and which would extract simultaneously the F_0 contours of several concurrent voices would have a very broad range of applications in computational auditory scene analysis (CASA), speech recognition, prosody analysis, speech enhancement or speaker identification. Several algorithms already exist that deal with the tracking of multiple F_0s (see for example [4, 5] and references therein), often relying on an initial frame-by-frame analysis followed by post-processing to reduce errors and obtain a smooth F_0 contour, for example using hidden Markov models (HMM) (see [4] for a review). Here we propose to perform estimation and model-based interpolation simultaneously, through a parametric model of the time and frequency shape of the spectral envelope of speech, based on a multi-pitch analysis method initially developed for feature extraction of music signals, the Harmonic-Temporal structured Clustering (HTC) method [6].

We will first give an overview of our method in Section 2, then show its accuracy and wide applicability through various experiments in Section 3.

This paper is a condensed version of the article [1] submitted for publication in IEEE Transactions on Audio, Speech and Language Processing.

2. FORMULATION OF THE MODEL

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 $\mu_k(t)$.

As described in [6], 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

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

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

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

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.

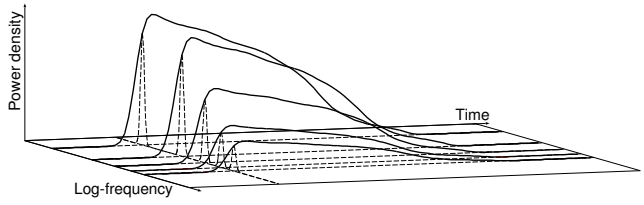


Fig. 1. Profile of a HTC source model $q_k(x, t; \mathbf{\Theta})$

2.2. Speech modeling

In the following, in order to model the spectrum of a speech utterance, we will 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 piece-wise 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, so as to be able to deal in the future with a wide variety of acoustic phenomenons (background music, phone ringing, etc). Moreover, their simple algebraical formulation is convenient for the optimization of the model.

The analysis interval is divided into subintervals $[t_i, t_{i+1}[$ which are assumed of equal length. The parameters of the spline contour model are then the values z_i of the F_0 at each bounding point t_i. The values z_i'' of the second derivative at those points are given by the expression $\boldsymbol{z}'' = M\boldsymbol{z}$ for a certain matrix M which can be explicitly computed offline, under the hypothesis that the first-order derivative is 0 at the bounds of the analysis interval. This can be assumed without loss of generality if we place the interval bounds outside the region where there is speech. One can then classically obtain a simple algebraic formula for the contour $\mu(t; \boldsymbol{z})$ on the whole interval. For $t \in [t_i, t_{i+1}[$:

$$\mu(t;\boldsymbol{z}) \triangleq \frac{1}{t_{i+1}-t_i}\Big(z_i(t_{i+1}-t) + z_{i+1}(t-t_i) \\ - \tfrac{1}{6}(t-t_i)(t_{i+1}-t)\big[(t_{i+2}-t)z_i'' + (t-t_{i-1})z_{i+1}''\big]\Big). \quad (3)$$

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. Each has a pattern of harmonic amplitudes that is fixed over time (this is expressed by assuming that $u_{kny} = u_{ky}, \forall n$), but the transitions between successive models allow the source to vary in spectral shape as well as F_0. 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. The analytical expression (3) of the spline F_0 contour is plugged into (2).

It is shown in [6] and [1] 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.

An example is presented in Fig. 2, based on the Japanese sentence "Tsuuyaku denwa kokusai kaigi jimukyoku desu" uttered by a female speaker. The F_0 contour estimated through our method is reproduced on both the observed and modeled spectrograms (after 30 iterations of the estimation algorithm) to show the accuracy of our algorithm.

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. Optimization of the noise model parameters is performed simultaneously with the optimization of the other parameters in the same EM-like framework [1].

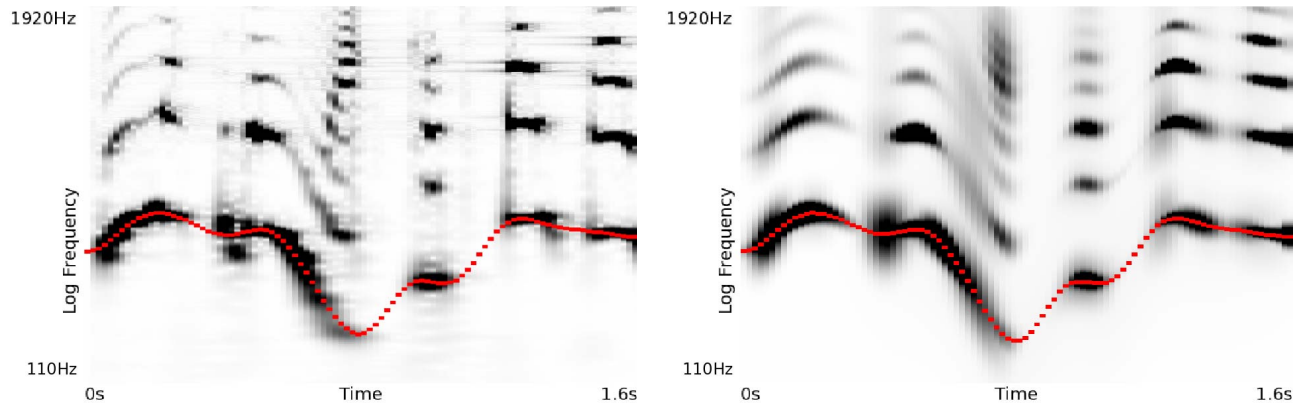
An estimation of the spectrogram where the noise has been canceled can be obtained from the original spectrogram by looking, at each point (x, t), at the proportion of the "clean" part inside the whole parametric model. In the course of the optimization of the model, it is on this "cleaned" part of the spectrogram that the F_0 contour estimation is performed, enabling our F_0 estimation algorithm to perform well even in very noisy environments, as we will show in the next section.

2.4. Parametric representation and potential applications

We would like to stress the fact that our algorithm not only estimates the F_0 contour, but also gives a parametric representation of the voiced parts of the spectrogram. This can be useful especially in the analysis of co-channel speech by multiple speakers, as one can get a parametric representation of the harmonic parts of the separated spectrograms of each utterance, which could be used to cluster the spectrogram of the mixed sound and separate the speakers, as well as for noise canceling, as we mentioned above.

3. EXPERIMENTAL EVALUATION

In the following experiments, for each input signal the power spectrum $W(x, t)$ was calculated using a Gabor wavelet transform. For the precise settings used in each of these experiments, we shall refer to [1].



(a) Observed spectrogram and estimated F_0 contour (b) Modeled spectrogram and F_0 contour

Fig. 2. Comparison of observed and modeled spectra ("Tsuuyaku denwa kokusai kaigi jimukyoku desu", female speaker).

3.1. Single-speaker F_0 estimation in clean environment

We evaluated the accuracy of the F_0 contour estimation of our model on a database of speech recorded together with a laryngograph signal [7], consisting of one male and one female speaker who each spoke 50 English sentences for a total of 0.12h of speech, for the purpose of evaluation of F_0-estimation algorithms. We used as ground truth the F_0 estimates and the reliability mask derived by de Cheveigné et al. [3]. Deviations over 20% from the reference were deemed to be gross errors.

The results can be seen in Table 1, with for comparison the results obtained by de Cheveigné et al. [3] for several other algorithms. References and details concerning these algorithms can be found in [3]. We can see that our model's accuracy for clean speech is comparable to the best existing single speaker F_0 extraction algorithms designed for that purpose.

3.2. Single F_0 estimation on speech mixed with white and pink noise

We performed F_0 estimation experiments on speech to which a white noise, band-passed between 50Hz and 3300Hz, was added, with SNRs of 0dB, -2dB and -10dB. The database mentioned above [7] was again used, and the white noise added was generated independently for each utterance. We also performed experiments with pink noise, band-passed between 50Hz and 3300Hz, with a SNR of -2dB. The spectrum of pink noise is closer to that of speech than white noise. The noise model was used.

As a comparison, we present results obtained on the same database using YIN [3] and the algorithm of Wu, Wang and Brown [5] (referred to as the WWB algorithm), specifically designed for F_0 tracking in a noisy environment, and that can also handle the estimation of two simultaneous F_0s. Although its parameters can be tuned on new databases to obtain the best performances, they are claimed to work fine in the version made available on the Internet (trained on a corpus [8] that we will use later), which we used for the experiments.

Table 1. *Gross error rates for several F_0 estimation algorithms on clean single speaker speech*

Method	pda	fxac	fxcep	ac	cc	shs	acf	nacf	additive	TEMPO	YIN	HTC
Gross error (%)	19.0	16.8	15.8	9.2	6.8	12.8	1.9	1.7	3.6	3.2	1.4	3.5

We obtained good results, presented in Table 2, showing the robustness of our method on noisy speech, when noise is not harmonic. YIN and the WWB algorithm were both outperformed, although we should note again that their code was used as is, whereas ours was developed with the task in mind. Thus, this comparison may not do them full justice.

3.3. Validation on a corpus of speech mixed with a wide range of interferences

In order to show the wide applicability of our method, we also performed experiments using a corpus of 100 mixtures of voiced speech and interference [8], commonly used in CASA research. The results we present for the WWB algorithm differ from the ones given in [5] as the criterion we use is different. To be able to compare it with our method, which does not perform a voiced-unvoiced decision, we do not take into account errors on the estimation of the number of F_0s, but only look at the accuracy of the output of the F_0 determination algorithm. Moreover, we focus on the F_0 estimation of the main voiced speech, as we want here to show that our algorithm robustly estimates the F_0 in a wide range of noisy environments. The ten interferences are grouped into three categories: 1) those with no pitch, 2) those with some pitch qualities, and 3) other speech. The reference F_0 contours for the ten voiced utterances were built using YIN on the clean speech and manually corrected.

The noise model was used for HTC, and the results are presented in Table 3. One can see that our algorithm again outperforms YIN and the WWB algorithm in all the interference categories.

Table 2. *Accuracy (%) of the F_0 estimation of single speaker speech mixed with white and pink noises*

	HTC (YIN,WWB)			
	White noise			Pink noise
	SNR=0dB	SNR=-2dB	SNR=-10dB	SNR=-2dB
Female speaker	95.8 (83.5, 56.1)	96.3 (77.8, 48.8)	88.2 (36.7, 09.0)	91.9 (46.1, 44.6)
Male speaker	92.1 (82.5, 69.3)	92.2 (77.2, 59.2)	79.7 (41.5, 19.2)	74.0 (58.1, 37.6)
Total	94.0 (83.0, 62.5)	94.3 (77.5, 53.8)	84.1 (39.0, 13.9)	83.2 (51.9, 41.2)

Table 3. *Accuracy (%) of the F_0 estimation of voiced speech with several kinds of interferences*

	HTC	WWB	YIN
Category 1	99.7	90.8	93.1
Category 2	98.6	96.1	75.7
Category 3	99.5	97.8	87.1

Table 4. *F_0 estimation of concurrent speech*

Gross error threshold	20%		10%	
	HTC	WWB	HTC	WWB
Male-Female	93.3	81.8	86.8	81.5
Male-Male	96.1	83.4	87.9	69.0
Female-Female	98.9	95.8	95.6	90.8
Total	96.1	87.0	90.2	83.5

3.4. Multi-pitch estimation

We present here results on the estimation of the F_0 contour of the co-channel speech of two speakers speaking simultaneously with equal average power. We used again the database mentioned above [7], and produced a total of 150 mixed utterances, 50 for each of the "male-male", "female-female" and "male-female" patterns. Our algorithm is used with two spline F_0 contours.

The evaluation was done in the following way: only times inside the reliability mask of either of the two references were counted; for each reference point, if either one of the two spline F_0 contours lied within a criterion distance of the reference, we considered the estimation correct. We present scores for two criterion thresholds: 10% and 20%. For comparison, tests using the WWB algorithm were also performed. Results summarized in Table 4 show that our algorithm outperforms the WWB algorithm on this experiment.

4. CONCLUSION AND FUTURE WORK

We introduced a new model describing the spectrum as a sequence of spectral cluster models governed by a common F_0 contour function, with smooth transitions in the temporal succession of the spectral structures. The model enables an accurate estimation of the F_0 contour on the whole utterance by taking advantage of its voiced parts in clean as well as noisy environments. We performed several experiments to evaluate the accuracy of our method. On single speaker clean speech, we obtained good results which we compared with existing methods specifically designed for that task. On co-channel concurrent speech, single speaker speech mixed with white noise, pink noise, and on a corpus of single speaker speech mixed with a variety of interfering sounds, we showed that our algorithm outperforms existing methods.

We are currently working on using the precise parametric expression and clustering of the spectrogram we obtained for noise canceling, speech enhancement and speech separation.

5. 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," *IEEE Trans. Audio Speech Language Process.*, submitted in 2006.

[2] W. J. Hess, *Pitch Determination of Speech Signals*, Springer, New York, 1983.

[3] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, 2002.

[4] A. de Cheveigné, "Multiple F_0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D.-L. Wang and G. J. Brown, Eds. IEEE Press / Wisley, 2006 (in press).

[5] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[6] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Speech Audio Process.*, 2006, in Press.

[7] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F_0 contours for computer and intonation teaching," in *Proc. Eurospeech*, 1993, pp. 1003–1006.

[8] M. P. Cooke, *Modeling Auditory Processing and Organisation*, Ph.D. thesis, University of Sheffield, 1993.