USING ENHANCED F0-TRAJECTORIES FOR MULTIPLE SPEAKER DETECTION IN AUDIO MONITORING SCENARIOS

Alessia Cornaggia-Urrigshardt, Frank Kurth

Fraunhofer FKIE, Communication Systems
Fraunhoferstr. 20
53343 Wachtberg, Germany

ABSTRACT

We propose to use enhanced F0-trajectories, which are extracted using shift-autocorrelation (shift-ACF), for multiple speaker detection in audio monitoring scenarios. After introducing spectral shift-ACF features, their performance in multiple F0-extraction in the presence of different noise types is estimated for synthetic signal scenarios. Afterwards, a novel method for F0-supertrajectory extraction is proposed and evaluated for multiple speaker detection in the presence of background noises that typically occur in audio monitoring. It turns out that due to their improved sharpness in representing harmonic components, spectral shift-ACF features outperform classical features in many cases.

Index Terms— Multiple Speaker Detection, Audio Monitoring, F0-Trajectories, Shift-ACF

1. INTRODUCTION

The detection of speaker activity is an important task in audio monitoring and acoustic scene analysis. This task can be particularly challenging due to various types of background noises and secondary signals present in a recorded signal. In this paper, we contribute to the potentially even more complex task of detecting the activity of multiple concurrently active speakers in a one-channel signal. More precisely, we are interested in estimating a list of all, possibly overlapping speech segments – each spoken by a single speaker – contained in a given signal. Using an additional speaker assignment, a subsequent application could be speaker diarization.

To separate two concurrently active speakers, a time-domain cancellation model was proposed in [1] two decades ago. In that work, while focusing on clean speech, harmonic properties and measurement of fundamental frequencies (F0) are used. Estimation of the number of speakers present in a segment of clean speech – which can be seen as a classification problem – was investigated based on the modulation spectrum [2], MFCC-feature-based GMM/HMMs and spectral peak extraction with subsequent clustering [3]. To incorporate temporal information, pitch-trajectories were used to model F0+harmonics in multiple speaker estimation [4] and extraction of polyphonic notes in music analysis [5].

While the previously reported approaches majorly deal with clean speech scenarios, to the best of our knowledge, only limited results dealing with monitoring scenarios are available. Based on the observation that tonal speech components are more robust to common noises and distortions than non-tonal ones, this paper aims at improving multiple speaker detection in noisy signals based on robustly extracted F0-trajectories. In particular, we propose to use enhanced F0-trajectories, which are obtained from the signal using the recently proposed shift-ACF [6]. In [7], it was shown that shift-ACF improves performance of F0-estimation compared to classical approaches particularly for noisy scenarios. A more detailed analysis on detection of sequences of temporal burst pulses [8] shows that shift-ACF is particularly robust against impulsive noises. While F0-estimation can be interpreted as a frequency-domain dual of the latter scenario, the time-varying structure of F0-trajectories makes it a somewhat different task. In this paper, we show that despite this difference F0-trajectories can be estimated robustly in the presence of mixtures of strong Gaussian and sinusoidal noises.

The paper is organized as follows. Sect. 2 briefly summarizes shift-ACF, spectral shift-ACF, and classical feature extractors used in this paper. The subsequent Sect. 3 presents our baseline experiment where performance of shift-ACF features for multiple speaker detection is systematically compared to classical features. In Sect. 4, a strategy for multiple speaker detection is proposed which is evaluated in Sect. 5.

2. SPECTRAL FEATURE EXTRACTION

The use of spectral features is motivated by observing that the fundamental frequency of voiced speech results in a high energy region within the short time spectrum x around a frequency F0 as well as at the harmonic frequencies 2·F0, 3·F0, ... Hence classical autocorrelation defined by $\mathrm{ACF}[x](s) := \sum_{k\in\mathbb{Z}} x(k)\cdot\overline{x(k-s)}$ as a typical mechanism for detecting repeating signal components is expected to show a local maximum at F0. As F0 is usually changing with time depending on speech prosody, we compute the shift-ACF for successive time frames of a speech signal y. To

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

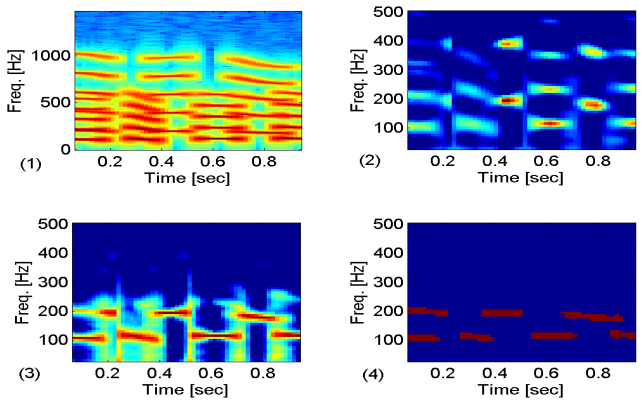

Fig. 1. Spectrogram of overlapping synthetic F0-trajectories with five harmonics each (1) as well as features obtained by column-wise classical ACF (2), a type 100 spectral shift ACF (3), and ground truth F0-trajectories (4).

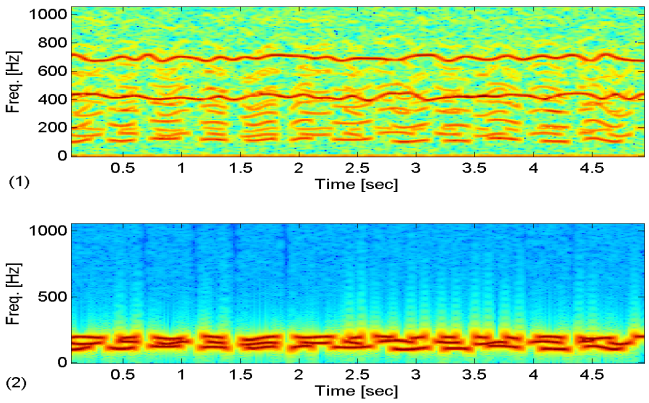

Fig. 2. (1) Five second sample for base mix of three overlapping synthetic speakers plus Gaussian noise at 0 dB and two random frequency-variant sinusoidal interferers. (2) Piecewise F0 trajectories used to generate synthetic speech in (1).

this end, we compute the spectrogram $\mathrm{SG}[y]$, where column j, $\mathrm{SG}[y]_{:,j}$, is obtained as discrete Fourier transform of a windowed version (125 ms Hann window, step size 16ms) of the j-th time frame $(y_{jS}, \ldots, y_{jS+N-1})$ of length N extracted from y using step size S. Then the *spectral ACF* is defined as $\mathrm{SpACF}[y](s,j) := \mathrm{ACF}[\mathrm{SG}[y]_{:,j}](s)$, i.e., by computing the ACF for all of the spectrogram columns individually.

Fig. 1 (1) shows the spectrogram of a one second length signal y containing a mix of two synthetic speakers as described in Sect. 3. In (2), $\mathrm{SpACF}[y]$ is shown, indicating the F0s of both speakers which are in the regions of 90–110 Hz (speaker 1) and 170-200 Hz (speaker 2). As documented by the ground truth (4) showing the true F0 as 7 *F0-trajectories*, the ACF in (2) does not properly represent all of those trajectories, which is due to closeness and partial overlap of harmonics. Moreover, the ACF contains considerable harmonic components that may complicate estimation of the true F0.

To overcome ACF-drawbacks, in [7] it was proposed to exploit *multiple* repetitions, i.e., that F0 usually has more than one harmonic component (in Fig. 1 (1) all of the shown F0-trajectories have four harmonics). This is achieved by first introducing both a *type 0* shift-product operator $\mathcal{O}_s^0 := x(k) \cdot \overline{x(k-s)}$ and a *type 1* shift-minimum operator $\mathcal{O}_s^1 := \min(|x(k)|, |x(k-s)|)$. Intuitively the shift-product, which is part of the above ACF-definition, can be used to amplify repeating components in x whereas the shift minimum can be used to suppress non-repeating components in x. Iterating those operators by $\mathcal{O}_s^t := \mathcal{O}_s^{t_1} \circ \cdots \circ \mathcal{O}_s^{t_n}$ with a length $n =: |t|$ *shift type* $t = (t_1, \ldots, t_n) \in \{0,1\}^n$ may then be used to enhance multiply repeating components. If we expect $n+1$-fold repetitions, using operators of length $\leq n$ turns out to be reasonable. The *shift-ACF of type* t is then defined as $\mathrm{ACF}^t[x](s) := \sum_{k\in\mathbb{Z}} \mathcal{O}_s^t[x](k)$. Note that ACF^0 is the classical ACF. As remarked in [6], multiply repeated components in x at lag s show up as peaks in shift-$\mathrm{ACF}^t(s)$ where the peak sharpness improves as a function of $|t|$.

Analogously to the case of ACF, now $\mathrm{SpACF}^t[y](s,j) := \mathrm{ACF}^t[\mathrm{SG}[y]_{:,j}](s)$ defines the *spectral shift-ACF* of type t. In our example, Fig. 1 (3) shows $\mathrm{SpACF}^{100}[y]$, where an improved sharpness of F0-trajectories and a better separation of both speakers as compared to the ACF in (2) can be observed. In the subsequent experiments, we will use the column-wise Fourier transform and the column-wise Cepstrum (both w.r.t. the columns of $\mathrm{SG}[y]$) as two additional, classical, (2D-) representations for F0-trajectory extraction.

3. BASELINE FEATURE EVALUATION

Prior to extracting F0-trajectories, we investigate how the different 2D features are able to represent F0-trajectories in a baseline experiment using synthetic speech. This allows us to both precisely control the ground truth of F0-trajectories and to decouple feature analysis from the analysis of the trajectory extraction method. In our experiments, we use $\mathrm{SpACF}^0 = \mathrm{SpACF}$, SpACF^{00}, SpACF^{100}, column-wise Fourier, and Cepstrum as 2D feature-representations. We also evaluated other shift-types up to length four, which in our test cases did not further improve performance.

To generate a *synthetic speaker* signal, we use a simplistic model of voiced speech. Each such signal is generated as a sequence of tonal components each consisting of a base F0-frequency and a number of harmonics. Component durations (between 120 and 400 ms) and pauses (between 80 and 150 ms) are randomly selected. To mimic prosody, F0 is allowed to vary by a certain amount (30 Hz in our tests) within a component. Variation is obtained by modulating F0 with a randomly generated spline function. Harmonics have energies which are linearly decaying to 25% of the F0. We critically remark that this is only a coarse speech model not including other parameters such as formant-based weighting

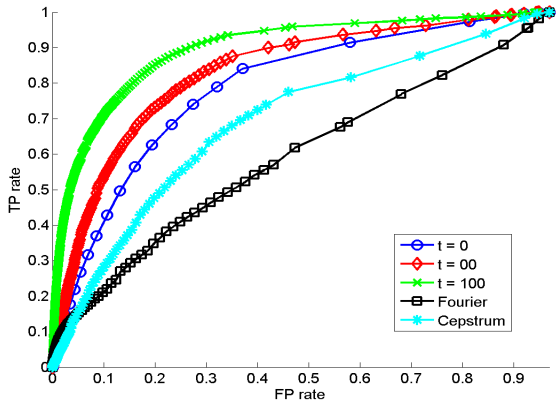

Fig. 3. ROC curve for base mix of three synthetic speakers plus Gaussian noise at 0 dB and 5 sinusoidal interferers.

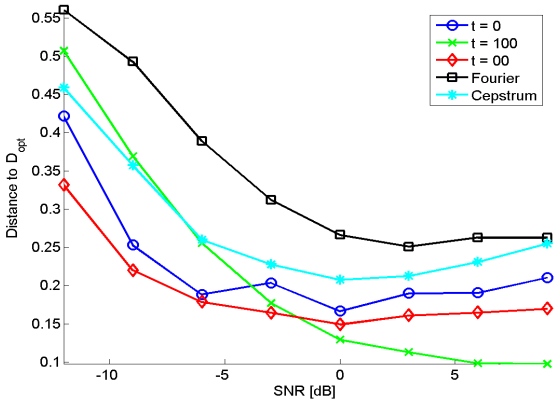

Fig. 4. Performance shown as distance to D_{opt} for added Gaussian noise (-12 to +6 dB) and three speaker base mix.

of the harmonics' energies or non-tonal components.

Based on M synthetic speaker signals s_i, a speech mix is then obtained as a weighted sum $s := \sum_{i=1}^{M} \alpha_i s_i$, where $0 \leq \alpha_i \leq 1$. To simulate background noise, Gaussian noise is added at a desired SNR. Structured interfering noise is modeled by adding a number of randomly modulated sinusoidals. As our analysis method works on short-time spectra, such interferers could be expected to have a strong impact on performance. Moreover, they model several types of realistic sounds such as produced by animals, sirens, or motors. Fig. 2 (1) shows a 5 second mix of three synthetic speakers (*base mix*) with base F0s at 100, 140, and 175 Hz, weighted using $(\alpha_1, \alpha_2, \alpha_3) = (1, 0.8, 0.5)$. Each tonal component consists of five harmonics only, which is realistic in several monitoring scenarios. Gaussian noise was added at 0 dB as well as two sinusoidal interferers, each having an energy equivalent to 5 times the F0-energy of the strongest speaker. For our evaluations, the underlying F0-trajectories, Fig. 2 (2), are used to generate ground truth F0-regions as illustrated in the introductory example, see Fig. 1 (4).

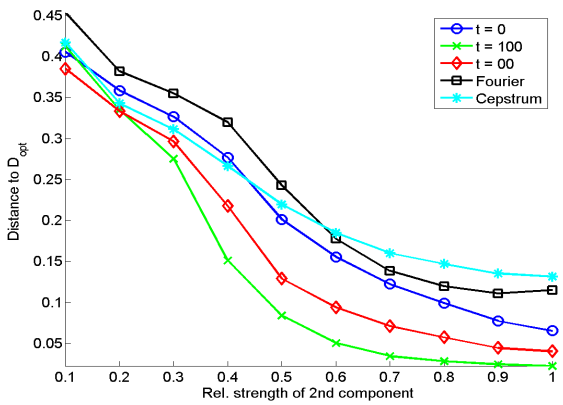

Fig. 5. Performance shown as distance to D_{opt} for two synthetic speakers with base F0 frequencies of 100 and 170 Hz, and relative strength $0.1 \leq \alpha \leq 1$ of speaker two.

In evaluating our (normalized) 2D feature-representations, we consider feature energy above a certain detection threshold. All positions with values above this threshold are counted as detections. Detections falling into ground truth F0 regions are correct detections, all others are counted as false alarms. The detection threshold is varied to obtain ROC-curves as depicted in Fig. 3, showing false positives (FP) rate versus detection, or true positive (TP), rate. Optimum performance is obtained at (TP,FP) $= (1, 0) =: D_{\text{opt}}$. Fig. 3 shows ROC-curves for the above base-mix of three speakers. We used an average over 50 signals of 5 seconds length each. Each signal contained Gaussian noise at 0 dB and five interfering sinusoids between 100 and 1000 Hz randomly choosen for each signal. Clearly, shift-ACF outperforms classical ACF and classical feature types. Robustness against various energy levels of Gaussian background noise is illustrated in Fig. 4. Each value on each curve is obtained from a ROC-evaluation and represents the mimimum distance of the respective ROC-curve from D_{opt}, i.e., small values are better than larger ones. We remark that we also estimated the common measure of equal error rate that yields very similar results. From Fig. 4 we see that type 100 shift-ACF performs best down to about -2dB, while for very low SNRs type 00 shift-ACF is better. The latter is due to the initial shift-minimum step in the type 100 operator suppressing some of the lower-energy harmonics which are close to the noise. Opposed to this, in the type 00 operator only harmonicity-enhancing shift-product steps are used. Note also the slightly decreasing performance for some of the features towards higher SNRs. This happens because feature energy that is widely distributed around the true F0 exceeds the noise floor in those cases and is thus counted as a (false) detection. In this sense, our evaluation favors concentrated feature energy. Feature performance for systematically varied speaker energies was investigated for a two speaker mix $s = s_1 + \alpha s_2$ with varying $0.1 \leq \alpha \leq 1$ and base F0s of 100 and 170 Hz at an SNR of 6 dB. While confirming

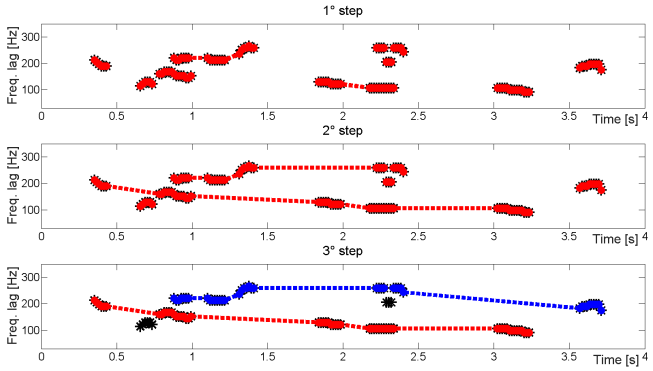

Fig. 6. Top to bottom: Three-step process of grouping F0-trajectories to F0-supertrajectories (dashed lines).

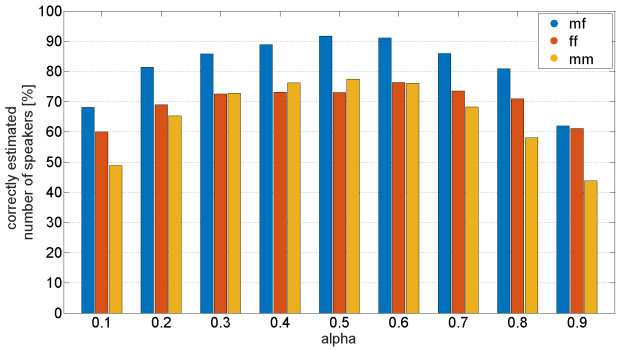

Fig. 7. Accuracy in detecting two speakers mixed at different weightings for combinations mf, ff, and mm.

superiority of shift-ACF features, the results shown in Fig. 5 indicate that there is a significant reduction in performance starting at about $\alpha = 0.45$ of speaker 2 energy.

4. F0-SUPERTRAJECTORY EXTRACTION

While the baseline experiment presented in Sect. 3 shows the features' potential by investigating feature concentration around F0, this section proposes to use estimated F0-trajectories for multiple speaker detection.

To extract basic F0-trajectories from a signal segment y, we process the spectral shift-ACF (e.g., Fig. 1 (3)). In particular, we first perform a column-wise peak picking on $\text{SpACF}^t[y]$. This is followed by concatenating adjacent peaks to paths using an optimization approach, see [7]. This results in short F0-trajectories as shown in Fig. 6 (top). In a subsequent step, we attempt to group F0-trajectories of different speakers. To this end, we assume that (1) temporally overlapping F0-trajectories belong to different speakers and (2) non-overlapping F0-trajectories which are close in time and frequency are likely to belong to the same speaker. Based on those assumptions, we extract *F0-supertrajectories* by iterating a sequence of grouping steps. Initializing F0-supertrajectories with all of the basic F0-trajectories, each subsequent iteration groups temporally successive (w.r.t. a threshold) F0-supertrajectories S_1 and S_2 where the frequency difference of either the closest points (end-point of S_1 and start-point of S_2) or the median F0s is below a threshold. In each iteration, both temporal and frequency thresholds are increased, thus successively allowing longer "gaps" to be filled. Fig. 6 shows three steps of this iteration where all new groupings are indicated by dashed lines. Resulting supertrajectories are represented by different colors. In the example, there are two supertrajectories corresponding to two speakers. As a result of this method, regions of speech activity can finally be derived from start- and end-points of the F0-supertrajectories. In our experiments, short supertrajectories of a duration below 0.5 seconds, were discarded in the end (black trajectories in Fig. 6).

5. EVALUATION

The performance of our speaker detection method was evaluated using a set of speech signals which were created under controlled conditions using spontaneous male (m) and female (f) speech from the German Kiel corpus[1].

In a first experiment, we evaluate detection of two speakers s_1, s_2 mixed using different intensities. Test signals for the three speaker combinations mf (45 minutes), ff (35 min.), and mm (35 min.) were generated using 4-second blocks each containing a weighted speech signal $\alpha s_1 + (1-\alpha)s_2$ for $0.1 \leq \alpha \leq 0.9$. As screening experiments indicated that speakers with base F0-frequencies closer than about 30 Hz are difficult to distinguish, we chose the speech mixes such that median F0 of both speakers in a 4-second block exceeds 30 Hz. Fig. 7 shows the accuracy in detecting the two speakers for varying values of α using a type 100 spectral shift-ACF.

In a second experiment, we tested the robustness against different noise types, in particular white Gaussian noise, street noise, lawnmower sounds, and thunderstorm recordings, which are added to speech signals at specific SNRs. The dataset consists of 45 minutes containing one speaker only and 45 minutes with two simultaneous speakers of equal energies. As noise-only signals, we used 18 minutes of street-, 4 minutes lawnmower- and 2 minutes thunderstorm-noise. Using more data of the – potentially more problematic – street noise can be motivated as, for many realistic scenarios, street noise is more common. In our tests we added these noises at SNRs of 0, 5, 10, 15, and 20 dB.

For SNRs of 0 dB and 15 dB, Table 1 shows confusion matrices for classifying the number of speakers based on our approach. The table shows results for street, lawnmower and thunderstorm background noise. Gaussian noise as well as the other SNR cases lead to qualitatively similar results and are

[1] http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html

(a)	**0**	**1**	**2**	**3**
0	**100**	0	0	0
1	1.2	**92.5**	6.1	0.2
2	1.3	33.2	**62**	3.5

(b)	**0**	**1**	**2**	**3**
0	**100**	0	0	0
1	0.3	**91.5**	8.2	0.2
2	0	0.4	**86**	10

(c)	**0**	**1**	**2**	**3**
0	**100**	0	0	0
1	1.5	**89.1**	9.4	0.2
2	1.2	26.8	**65.6**	6.4

(d)	**0**	**1**	**2**	**3**
0	**100**	0	0	0
1	0.4	**89.5**	10	0.1
2	0	3	**87**	10

(e)	**0**	**1**	**2**	**3**
0	**100**	0	0	0
1	1.5	**92.7**	5.8	0
2	0.9	28.9	**66.4**	3.8

(f)	**0**	**1**	**2**	**3**
0	**100**	0	0	0
1	0.3	**92.1**	7.3	0.3
2	0	2.8	**84.3**	12.9

Table 1. Confusion matrices: ground truth (rows; 0–2 speakers) vs number of speakers returned by algorithm (columns; 0–3 speakers) shown for added backgrounds of street noise at SNRs of (a) 0 dB and (b) 15 dB, lawnmower noise at SNRs of (c) 0 dB and (d) 15 dB, thunderstorm noise at SNRs of (e) 0 dB and (f) 15 dB.

not shown here because of space restrictions. The results indicate that no false positives were returned for recordings with background noise only. In particular, even in cases where trajectories where present, these were discarded due to shortness of resulting supertrajectories. The accuracy for detecting two speakers even in case of significant noise (SNR = 0 dB) is around 65% and it exceeds 85% at SNRs over 15 dB. Furthermore, with increasing SNR we see that errors are mostly due to false positive (i.e., 3 speakers estimated instead of 2), while for low SNR values we have mostly false negatives (i.e., 0 or 1 speakers estimated instead of 2).

In a third experiment, we compare performance before and after extracting F0-supertrajectories. To this end, we added speech signals from the Kiel Corpus at random positions of signals containing street background noise (using an SNR of 10 dB), resulting in a total of 21 minutes of test material (6.4 minutes noise only, 6.4 minutes one speaker only, 8.2 minutes containing two speakers). Evaluation of automatic F0-trajectory extraction was then performed by comparison to manually annotated F0-ground truth, where extracted F0-trajectories in an ε-neighborhood of the ground truth were counted as correctly detected, and as false positives otherwise. Extracted supertrajectories are evaluated in the same way, taking into consideration the trajectories that compose them. Here, trajectories are considered to be false positives if they are included in the supertrajectory of a wrong speaker. Choosing $\varepsilon = 15$ Hz, we obtain a performance of $(\mathrm{TP}, \mathrm{FP}) = (0.82, 0.14)$ for trajectory extraction only and $(\mathrm{TP}, \mathrm{FP}) = (0.79, 0.11)$ after supertrajectory extraction. The lower FP-rate in the latter is due to trajectories coming from non-speech components being discarded in constructing supertrajectories. The lower TP-rate indicates trajectories being assigned to the wrong speaker.

6. CONCLUSIONS

This paper presents two contributions to multiple speaker detection in audio monitoring recordings. Firstly, we propose to use spectral shift-ACF features to obtain a robust representation of F0-trajectories. From a comprehensive baseline experiment using synthetic speaker signals, we conclude that the improved energy concentration of shift-ACF features around F0 makes them robust to both strong Gaussian and sinusoidal interfering noises. Moreover, improved performance in representing multiple speaker F0s mixed at various energies has been shown. Secondly, we propose a method for multiple speaker detection based on extracting F0-supertrajectories. Experimental results on speech mixes and realistic background noise indicate that the method is already suitable for audio monitoring scenarios including up to two simultaneous speakers. Suitability of the proposed features for more than two speakers is indicated by experiments on synthetic signals. A promising direction for future work would be to combine the proposed method of using enhanced F0- (super-) trajectories with existing methods such as [4, 5] for estimating multiple mixed harmonic components.

REFERENCES

[1] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *JASA*, vol. 93, pp. 3271–3290, 1993.

[2] T. Arai, "Estimating Number of Speakers by the Modulation Characteristics of Speech," in *Proc. ICASSP*, 2003, vol. 2, pp. 197–200.

[3] Umer Rafi and Rolf Bardeli, "Harmonic Cues for Number of Simultaneous Speakers Estimation," in *AES 53rd Conference: Semantic Audio*, Jan. 2014.

[4] Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama, "Multi-Pitch Trajectory Estimation of Concurrent Speech Based on Harmonic GMM and Nonlinear Kalman Filtering," in *Proc. ICSLP*, 2004, vol. 1, pp. 2433–2436.

[5] Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 982–994, 2007.

[6] F. Kurth, "The Shift-ACF: Detecting Multiply Repeated Signal Components," in *Proc. IEEE WASPAA*, 2013.

[7] F. Kurth, A. Cornaggia-Urrigshardt, and S. Urrigshardt, "Robust F0 Estimation in Noisy Speech Signals Using Shift Autocorrelation," in *Proc. IEEE ICASSP*, 2014.

[8] Paul M. Baggenstoss and Frank Kurth, "Comparing Shift-ACF with Cepstrum for Detection of Burst Pulses in Impulsive Noise," *Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1574–82, 2014.