EVALUATION OF PITCH ESTIMATION ALGORITHMS ON SEPARATED SPEECH

Feng Huang[1,2]*, Yu Ting Yeung*[1]*, and Tan Lee*[1,2]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

[2]Shenzhen Research Institude, The Chinese University of Hong Kong, Shenzhen, China

{fhuang,ytyeung,tanlee}@ee.cuhk.edu.hk

ABSTRACT

To post-process outputs of speech separation systems with harmonic enhancement, it is normally required to estimate the fundamental frequency. This paper evaluates the performance of a few representative robust pitch estimation algorithms on speech reconstructed from two-speaker mixture signals. The separation outputs obtained by two state-of-the-art single-channel separation algorithms are used for the evaluation. A recently proposed sparsity-based pitch estimation method is applied to the separated speech and a new pitch tracking algorithm is proposed. Experimental results show that on the separated speech the proposed method consistently surpasses the others with significantly low gross error rate, which is similar to the gross error rates of the other methods on clean speech.

Index Terms— Robust pitch estimation, harmonic noise, speech separation, harmonic enhancement

1. INTRODUCTION

Automatic detection of fundamental frequency (i.e., pitch or F0) [1] from acoustic signal is fundamental for numerous applications of speech signal processing, e.g., speech coding, speech and speaker recognition, and harmonic enhancement (HE) [2, 3]. HE refers to the process of enhancing or restoring harmonic components of a speech signal and suppressing the non-harmonic ones. It is an effective approach to improving the quality of noise-corrupted speech as well as processed speech outputted by enhancement systems [2, 3, 4, 5]. In our previous study of speech separation [6], it was also suggested that performing HE on the separation output could be beneficial for improving speech quality. For effective HE, it is important to have correct estimation of the pitch.

To improve pitch estimation accuracy is one major research focus. There have been many robust algorithms developed [7, 8, 9, 10, 11, 12]. Most of the algorithms were designed for general additive noise such as white noise, speech-spectrum-shaped noise or car noise. These types of noise are distinguishable from speech as they do not have harmonic structure. There were few studies on applying pitch estimation algorithms to noise that has harmonic structure.

In this study, we investigate the performance of several representative methods of robust pitch estimation on separated speech. In speech separation, a mixture of speech signals is given and a separation algorithm is used for reconstructing the speech of a target speaker. In the reconstructed speech, the residual noise generally exhibits harmonic structure, which comes from the interfering speech source. Factorial hidden Markov models based algorithm [13] is one of the state-of-the-art separation methods. It uses acoustic models of the log spectrum of the sources as prior information for separation. It was observed that in the separated output there usually resided harmonic information from the interfering source (e.g. phase information [14]).

To deal with harmonic noise in the separated speech, we propose to apply a sparsity-based pitch estimation algorithm that we recently proposed in [15]. This algorithm utilizes prior speech information and l_1-regularized maximum likelihood (ML) estimation to improve estimation accuracy. We derive a localized formulation of this algorithm for separated speech and propose a pitch tracking algorithm for determining pitch track of the target speaker. The proposed approach with four other robust algorithms are evaluated and compared on speech separation outputs obtained by two recently developed single-channel separation algorithms.

2. ROBUST PITCH ESTIMATION ALGORITHMS

Pitch estimation is to detect the fundamental frequency of a quasi-periodic audio signal. The goal of noise-robust pitch estimation is to determine the fundamental frequency of underlying speech from a distorted observation.

2.1. Baseline approaches

CEP: Cepstrum is defined as the inverse Fourier transform of the log magnitude spectrum. Cepstrum of voiced speech has a sharp peak that corresponds to the fundamental frequency. The standing-out of this peak was found to be robust to noise [16]. By locating the strongest cepstral peak, the fundamental frequency is determined [17].

RAPT: The robust algorithm for pitch tracking (RAPT) [18] is a widely used time-domain approach. It employs two-pass normalized-cross correlation function, at reduced and original sampling rate, to reduce computational load and obtain a set of F0 candidates. Dynamic programming is used to determine the best pitch track.

PEFAC: The algorithm of pitch estimation filter with amplitude compression (PEFAC) [11] is a recently proposed frequency-domain method. It uses amplitude compression to

This work was supported in part by the General Research Funds (Ref: CUHK 414010 & CUHK 413811) from the Hong Kong Research Grants Council, and the Shenzhen Basic Research Grant (No. JCYJ20120831091405139).

attenuate narrowband noise, and then applies comb-filtering in log-frequency power spectral domain to sum the energy of hypothesized pitch harmonics. Pitch candidate that has the largest accumulated energy is selected. The filter can also combat noise that has smoothly varying power spectrum.

TAPS-AutoC: As described in [10], the temporally accumulated peak spectrum (TAPS) is computed by accumulating spectral peaks over consecutive frames. After accumulation, speech harmonic peaks are concentrated around the fundamental frequency and its multiples, while spectral peaks of noise tend to be irregularly located with relatively small amplitudes. Autocorrelation of TAPS is used to determine the frequency separation between the harmonic peaks and give an estimation of the fundamental frequency.

3. SPARSITY-BASED PITCH TRACKING

Based on TAPS, a sparsity-based framework was proposed for pitch estimation by employing prior information and sparse reconstruction techniques [15, 19]. In this section, we first briefly review the algorithm **TAPS-l_1-ML** [15] and derive a localized formulation for separated speech. A pitch tracking algorithm is then proposed to determine pitch track of the target speaker.

3.1. Sparsity-based pitch estimation (TAPS-l_1-ML)

Let $\mathbf{p} \in \mathcal{R}^{M\times 1}$ denote a peak spectrum, obtained by retaining peaks of the magnitude spectrum and setting the others to 0. Let $\mathbf{y}$ denote a TAPS, obtained by summing $\mathbf{p}$ of several, say K consecutive frames. Given an over-complete set of clean peak spectrum exemplars $\mathbf{A} = [\bar{\mathbf{p}}_1\ \bar{\mathbf{p}}_2\ \cdots\ \bar{\mathbf{p}}_n \cdots\ \bar{\mathbf{p}}_N] \in \mathcal{R}^{M\times N}$, with $N \gg M$, $\mathbf{y}$ is modeled as a sparse linear combination of the exemplars

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}, \tag{1}$$

where $\mathbf{x} \in \mathcal{R}^{N\times 1}$ is a weight vector assumed K-sparse (at most K nonzero elements), while $\mathbf{v} \in \mathcal{R}^{M\times 1}$ represents noise in the peak spectrum domain.

Given $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the sparse weight $\mathbf{x}$ is estimated by the following l_1-regularized minimization [15]:

$$\begin{aligned} \min_{\mathbf{x}} \quad & (\mathbf{A}\mathbf{x} - \mathbf{y} + \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{A}\mathbf{x} - \mathbf{y} + \boldsymbol{\mu}) \\ \text{subject to} \quad & \|\mathbf{x}\|_1 \le K \text{ and } \mathbf{x} > \mathbf{0}. \end{aligned} \tag{2}$$

Each non-zero element of the estimated weight $\hat{\mathbf{x}}$ corresponds to a constituent exemplar, which indicates a candidate pitch value. Hence, a set of pitch candidates $\{{f_0}^{\mathrm{c}}(q)|1 \le q \le Q\}$ is obtained with their corresponding weights $\{\hat{x}^{\mathrm{c}}_q|1 \le q \le Q\}$ available [15, 20]. For a pitch candidate ${f_0}^{\mathrm{c}}(q)$, a confidence measure [20] is computed as

$$\mathrm{P}_{\mathrm{F0}}(q) = \frac{\hat{x}^{\mathrm{c}}_q}{\sum_{q=1}^{Q} \hat{x}^{\mathrm{c}}_q}. \tag{3}$$

The candidate that has the largest confidence measurement is selected as the estimated pitch, i.e., $\hat{f}_0 = f_0^{\mathrm{c}}(q^*)$ with $q^* = \arg\max_q(\mathrm{P}_{\mathrm{F0}}(q))$.

This approach utilizes prior information of both speech (i.e., $\mathbf{A}$) and noise (i.e., $\boldsymbol{\mu}, \boldsymbol{\Sigma}$) for pitch estimation. It was shown to have good performance even at very low SNRs (e.g. < -5 dB) [15, 19].

3.2. Localized TAPS-l_1-ML for separated speech

Since residual noise in separated speech usually contains a small portion of the interfering source, $\mathbf{v}$ as in Eq. (1) may exhibit some harmonic components of the interfering speaker. It is nontrivial and could even be very difficult to learn the Gaussian parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for $\mathbf{v}$ in practice. Here we consider that $\mathbf{v}$ can be represented by the following sparse combination of another set of exemplars $\mathbf{A}_v$,

$$\mathbf{v} = \mathbf{A}_v \mathbf{x}_v + \mathbf{n}, \tag{4}$$

with $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Combining Eq. (1) and (4), we obtain

$$\mathbf{y} = \left[\ \mathbf{A}\ \mathbf{A}_v\ \right] \left[\begin{array}{c} \mathbf{x} \\ \mathbf{x}_v \end{array}\right] + \mathbf{n}.$$

Denote the above concatenated exemplar matrix and sparse weight vector as $\mathbf{A}^\dagger$ and $\mathbf{x}^\dagger$, respectively. We obtain

$$\mathbf{y} = \mathbf{A}^\dagger \mathbf{x}^\dagger + \mathbf{n}. \tag{5}$$

With the same principle as in Section 3.1, $\mathbf{x}^\dagger$ is estimated by

$$\begin{aligned} \min_{\mathbf{x}^\dagger} \quad & \|\mathbf{y} - \mathbf{A}^\dagger \mathbf{x}^\dagger\|_2^2 \\ \text{subject to} \quad & \|\mathbf{x}^\dagger\|_1 \le K + K_v \text{ and } \mathbf{x}^\dagger > \mathbf{0}, \end{aligned} \tag{6}$$

where K_v reflects the sparseness of $\mathbf{x}_v$.

The use of (6), instead of (2), avoids the difficulty of learning $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, since obtaining $\mathbf{A}^\dagger$ could be easier. The basic requirement [15] for constructing $\mathbf{A}^\dagger$ is that it should be complete enough to resolve the harmonic structure in $\mathbf{y}$, including that of the target speaker as well as that of the interfering speaker. Hence $\mathbf{A}^\dagger$ can be constructed by recruiting peak spectrum exemplars from both speakers. However, employing (6) on the other hand increases the difficulty of determining the correct pitch for the target speaker, because now $\mathbf{x}^\dagger$ also reveals harmonic components of the interfering speaker. In the following, we propose a pitch tracking algorithm for the **TAPS-l_1-ML** method, with the goal of producing a correct pitch track for the target speaker.

3.3. Pitch tracking algorithm for TAPS-l_1-ML

There are a number of pitch candidates at each frame. The purpose of pitch tracking is to find the best path of candidates across all frames. Dynamic programming (DP) is used here. With DP, the best path is defined as the one with highest score. The score for ${f_0}^{\mathrm{c}}(q)|_r$, i.e., the qth pitch candidate at the rth frame, is defined as

$$\begin{aligned} \mathrm{SCR}(q, r) = \mathrm{SCR}_{\mathrm{H}}(q, r) + \\ \mathrm{SCR}_{\mathrm{CF0}}(q, r, r-1) + \mathrm{SCR}_{\mathrm{CMF0}}(q, r, N). \end{aligned} \tag{7}$$

$\mathrm{SCR}_{\mathrm{H}}(q, r)$ gives a score for the harmonicity of the hypothesized pitch ${f_0}^{\mathrm{c}}(q)|_r$. Since the confidence measure defined in Eq. (3) reflects strength of the harmonic components of the candidate [20], it is used as a harmonicity index. $\mathrm{SCR}_{\mathrm{H}}(q, r)$ is computed by

$$\mathrm{SCR}_{\mathrm{H}}(q, r) = \mathrm{P}_{\mathrm{F0}}(q)|_r, \tag{8}$$

where $|_r$ indicates that it is computed for the rth frame.

$\mathrm{SCR}_{\mathrm{CF0}}(q, r, r-1)$ gives a score for the optimal path that connects ${f_0}^{\mathrm{c}}(q)|_r$ to frame $r-1$. When determining a pitch track, continuity of the path is a major consideration. In [3], the probability of percentage pitch change between neighboring frames, i.e., $\mathrm{P}_\delta(\frac{\delta f_0}{f_0})$, was analyzed. It was shown that large pitch change was less likely to happen. It is thus preferred to connect ${f_0}^{\mathrm{c}}(q)|_r$ to the previous-frame candidate whose pitch value is the closest. We further take into account the scores $\mathrm{SCR}(\cdot, r-1)$ of the previous frame, and define

$$\mathrm{SCR}_{\mathrm{CF0}}(q, r, r-1) = \max_l \left\{ \mathrm{P}_\delta \left(\frac{{f_0}^{\mathrm{c}}(q)|_r - {f_0}^{\mathrm{c}}(l)|_{r-1}}{{f_0}^{\mathrm{c}}(q)|_r} \right) \cdot \mathrm{SCR}(l, r-1) \right\}, \quad (9)$$

where $\mathrm{P}_\delta(\frac{\delta f_0}{f_0})$ follows the the empirical one described in [3]. The term $\mathrm{P}_\delta(\cdot) \cdot \mathrm{SCR}(l, r-1)$ computes a score for connecting ${f_0}^{\mathrm{c}}(q)|_r$ with ${f_0}^{\mathrm{c}}(l)|_{r-1}$. The $\max(\cdot)$ operation implies the selection of an optimal backward path for the candidate ${f_0}^{\mathrm{c}}(q)|_r$.

$\mathrm{SCR}_{\mathrm{CMF0}}(q, r, N)$ gives a score for encouraging consistence between ${f_0}^{\mathrm{c}}(q)|_r$ and F0 mean of previous voiced segments. We define

$$\mathrm{SCR}_{\mathrm{CMF0}}(q, r, N) = \mathrm{P}_\delta \left(\frac{{f_0}^{\mathrm{c}}(q)|_r - \bar{f}_0|_{r-1}^N}{{f_0}^{\mathrm{c}}(q)|_r} \right), \quad (10)$$

where $\bar{f}_0|_{r-1}^N$ is the mean of reliable F0s estimated in the past N frames, covering at lease one previous voiced segment. If a candidate pitch is selected as the estimated one, it will be marked as reliable given that its harmonicity score $\mathrm{SCR}_{\mathrm{H}}(\cdot, \cdot)$ is larger than a threshold [20].

In summary, the proposed score $\mathrm{SCR}(\cdot, \cdot)$ favors pitch candidates of strong harmonicity, continuity of pitch track and consistence of F0 mean in consecutive voiced segments.

For off-line processing, the scores are first computed in a forward manner, i.e., from the first frame to the last frame. Then by backward tracing, a global optimal pitch track is determined. Here we consider on-line processing scenario, where information of future frame is not available. For such case, pitch for the current frame r is determined as $\hat{f}_0|_r = f_0^{\mathrm{c}}(q^*)|_r$ with $q^* = \arg\max_q(\mathrm{SCR}(q, r))$.

4. EXPERIMENTS

4.1. Experimental setup

A set of speech data taken from the GRID corpus [21] is used for evaluation. The data set consists of utterances from 6 speakers, namely T1 (Male), T2 (Male), T17(Male), T18 (Female), T24 (Female) and T25 (Female). There are 25 utterances for each speaker. The utterances were mixed in the following way: T1+T2 (Male+Male), T17+T18 (Male+Female), and T24+T25 (Female+Female) with 0 dB signal-to-signal ratio. After mixing, there were 625 mixture signals for each target speaker. The mixture signals were then processed by two single-channel speech separation algorithms, i.e., factorial HMM based algorithm (FHMM) [13] and dynamic conditional random fields based algorithm (DCRF) [22]. The goal was to reconstruct the speech of the target speaker from the mixture signal. Details for the training of these two separation algorithms can be found in [22]. We used two parameter settings for both FHMM and DCRF, namely 16as (i.e., 16 acoustic states) and 512as (i.e., 512 acoustic states). Table 1 shows the objective quality measurements of the separated speech. The full set of separated signals[1] (625 utterances per speaker per separation condition) were re-sampled at 8 kHz. They are used for the evaluation of pitch estimation algorithms.

Table 1: SNR and PESQ of the separated speech signals

	$\mathrm{FHMM}_{16\mathrm{as}}$	$\mathrm{DCRF}_{16\mathrm{as}}$	$\mathrm{FHMM}_{512\mathrm{as}}$	$\mathrm{DCRF}_{512\mathrm{as}}$
SNR(dB)	3.75	5.37	6.59	7.73
PESQ	1.72	2.11	2.39	2.60

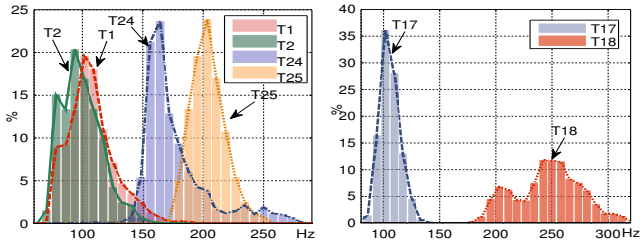

Fig. 1: Fundamental frequency distribution of the speakers

Pitch estimation accuracy is evaluated in terms of *gross pitch error* (GPE) and *fine pitch error* (FPE) [23]. An estimated $\hat{f}_0$ value is regarded as a GPE if $\hat{f}_0 \notin \left[f_{0,\mathrm{tar}}^{\mathrm{true}} - \epsilon,\ f_{0,\mathrm{tar}}^{\mathrm{true}} + \epsilon\right]$. Otherwise, it is regarded as an FPE. $f_{0,\mathrm{tar}}^{\mathrm{true}}$ is the reference pitch of the target speaker. $\epsilon = 16$ Hz. For GPE, the error rate in percentage is calculated. For FPE, the root mean square (RMS) of the deviation $|\hat{f}_0 - f_{0,\mathrm{tar}}^{\mathrm{true}}|$ is computed. Reference pitch and voicing status for result verification were first obtained by using the software Wavesurfer [24] and then manually verified. Distribution of the fundamental frequency of the speakers is shown in Fig. 1.

CEP was implemented following the algorithm in [17]. For **RAPT** and **PEFAC**, their implementations in Voicebox [25] are used. Frame size and frame shift are 60 ms and 12 ms, respectively. For **TAPS**-based algorithms, the number of accumulated peak spectrum K is set to 4 [10]. Peak spectrum for computing TAPS is obtained every 12 ms for a signal segment of 24 ms. Hence the TAPS-based methods also estimate a pitch value from a signal segment of 60 ms. FFT size is set to 1024. The dimension of peak spectrum vector is $M = 102$, which covers the frequency range from 0 Hz to 800 Hz.

For **TAPS-l_1-ML**, $150(25 \times 6)$ utterances, from the same group of speakers and not overlapping with the testing data, were used to train the prior information matrix $\mathbf{A}^\dagger$ [15]. There were 100 exemplars obtained for each speaker. These exemplars were concatenated to form an exemplar matrix $\mathbf{A}^\dagger \in \mathcal{R}^{600\times102}$. Libqp [26] is used to solve the quadratic programming problem (6), where K_v is empirically set to 1. For the tracking algorithm, N for Eq. (10) is set to cover a signal duration of 500 ms and the threshold of $\mathrm{SCR}_{\mathrm{H}}(\cdot, \cdot)$ for determining a reliable estimation is set to 0.6.

[1] Samples available online: www.ee.cuhk.edu.hk/~ytyeung/dmmse.htm

Table 2: GPE and FPE results on clean and mixed speech

		CEP	RAPT	PEFAC	TAPS-AutoC	TAPS-l_1-ML
Clean	GPE	8.7%	5.3%	4.7%	3.0%	**1.4%**
	FPE	5.32Hz	**3.02Hz**	3.50Hz	5.39Hz	3.21Hz
Mixed	GPE	43.2%	48.5%	43.7%	45.5%	**37.8%**
	FPE	6.39Hz	5.67Hz	**5.35Hz**	6.59Hz	5.90Hz

Table 3: DGPE rate (%) for VV and VVD of mixed speech

	CEP	RAPT	PEFAC	TAPS-AutoC	TAPS-l_1-ML
VV	9.0	21.4	10.7	15.1	**3.7**
VVD	10.1	30.3	15.3	22.3	**5.9**

4.2. Results on clean and mixed speech

Table 2 shows the results of the pitch estimation algorithms on clean and mixed speech. It can be seen that TAPS-l_1-ML has the lowest GPE rates. FPE results of RAPT, PEFAC and TAPS-l_1-ML are closely low.

For mixed speech, there were 625 mixture signals for each speaker. For a speaker, the estimated pitch tracks of the mixture signals were compared to the speaker's reference pitch to compute the results. Since a mixture signal was associated with two speakers, an estimated pitch track was compared twice for error counting. To further evaluate the performance of the algorithms on mixed speech, we compute another type of error rate for the following two kinds of signal slots:

VV Time slots where signal components of both speakers are all voiced;

VVD Time slots where signal components of both speakers are all voiced and the fundamental frequencies of both speakers are different ($> 2\epsilon$).

In VV, fundamental frequencies of both speakers may be the same. For instance, this situation is very likely to happen in the T1+T2 case (see Fig. 1). In VVD, fundamental frequencies of the speakers are distinct.

For mixed speech, an estimated pitch is regarded as a dual gross pitch error (DGPE) if it does not hit either one of the true pitch values (with $\pm\epsilon$ range). DGPE rate indicates how well a pitch estimation algorithm can perform in identifying either one of the two fundamental frequencies in a voiced-voiced mixture signal. Table 3 gives the DGPE rates of the evaluated algorithms. It can be seen that DGPE rates of TAPS-l_1-ML are significantly lower than the others for both VV and VVD.

4.3. Results on separated speech

Table 4 gives the GPE results for pitch estimation on the separated signals. Overall GPE rates computed from all voiced frames as well as GPE rates computed from VV and VVD frames are given. It can be observed that for both separation algorithms, GPE rates of all pitch estimation algorithms decrease when the number of acoustic state increases. This agrees with the results in Table 1 that the quality of separated speech is improved with more acoustic states. It is also shown that TAPS-l_1-ML consistently surpasses the other methods for all conditions. For the instance of $\mathrm{FHMM}_{512\mathrm{as}}$, GPE rate of TAPS-$l_1$-ML is 4.4%, which is lower than the GPE rate of PEFAC on clean speech, i.e., 4.7%.

Table 4: GPE rate (%) on separated speech

		CEP	RAPT	PEFAC	TAPS-AutoC	TAPS-l_1-ML
$\mathrm{FHMM}_{16\mathrm{as}}$		24.6	17.2	16.5	14.3	**7.3**
	VV	30.4	22.9	20.9	19.9	**10.4**
	VVD	35.7	25.9	23.0	22.7	**12.0**
$\mathrm{DCRF}_{16\mathrm{as}}$		21.1	14.3	11.5	11.7	**6.0**
	VV	26.1	19.1	15.0	16.2	**8.5**
	VVD	29.5	21.1	16.0	18.1	**9.5**
$\mathrm{FHMM}_{512\mathrm{as}}$		16.7	11.8	10.2	8.8	**4.4**
	VV	19.4	15.0	12.7	11.9	**6.1**
	VVD	20.2	14.9	11.0	11.5	**6.1**
$\mathrm{DCRF}_{512\mathrm{as}}$		14.9	10.0	7.8	7.4	**3.9**
	VV	17.0	12.3	9.6	9.9	**5.4**
	VVD	16.9	11.4	8.1	8.9	**5.0**

As for FPE, results at different separation conditions have a similar trend as that of the clean speech case shown in Table 2. FPE of RAPT tends to be the lowest, while FPE of PEFAC and TAPS-l_1-ML are both close to the lowest. CEP and TAPS-AutoC generally have the largest FPE. For the instance of $\mathrm{FHMM}_{512\mathrm{as}}$, the FPE results (in Hz) for RAPT, PEFAC, TAPS-l_1-ML, TAPS-AutoC and CEP are 3.80, 3.98, 4.28, 5.63 and 5.81, respectively.

5. CONCLUSION AND RELATION TO PRIOR WORK

This paper evaluated four representative and a new pitch estimation algorithms on speech reconstructed from two-speaker mixture signals. The sparsity-based algorithm, namely TAPS-l_1-ML, was applied to the separated speech. Based on TAPS-l_1-ML, a localized formulation was derived and a pitch tracking algorithm was proposed. Experimental results confirmed that the proposed algorithm consistently surpassed the other algorithms with significantly low gross error rate. Gross error rate of the proposed algorithm on the separated speech was similar to the gross error rates of the others on clean speech. With the fundamental frequency more accurately estimated, effective harmonic enhancement can be achieved to further improve the quality of separated speech.

[Relation to prior work] This study is an extension of our prior research on robust pitch estimation [10, 15, 19, 20]. It extends the application scope of the sparsity-based pitch estimation method [15, 20] and verifies its ability for coping with noise that has harmonic structure. It connects the previous study on speech separation [6, 22] and harmonic enhancement [5, 27], as it is our first step in developing a harmonic enhancement algorithm for post-processing of separated speech.

6. REFERENCES

[1] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer, Berlin, 1983.

[2] A.-T. Yu and H.-C. Wang, "New speech harmonic structure measure and it application to post speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2004*, May 2004, vol. 1, pp. 729–732.

[3] M. R. Every and P. J. B. Jackson, "Enhancement of harmonic content of speech based on a dynamic programming pitch tracking algorithm," in *Proc. Interspeech 2006*, 2006.

[4] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2005*, 2005, vol. 1, pp. 157–160.

[5] F. Huang, T. Lee, and W. B. Kleijn, "Transform-domain Wiener filter for speech periodicity enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2012*, Mar. 2012, pp. 4577–4580.

[6] Y. T. Yeung, T. Lee, and C.-C. Leung, "Integrating multiple observations for model-based single-microphone speech separation with conditional random fields," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2012*, 2012, pp. 257–260.

[7] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech, Audio Process.*, vol. 9, pp. 727–730, Oct. 2001.

[8] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proc. Interspeech 2007*, Aug. 2007, pp. 2765–2768.

[9] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "A pitch extraction algorithm in noise based on temporal and spectral representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2008*, 2008, pp. 4477–4480.

[10] F. Huang and T. Lee, "Pitch estimation in noisy speech based on temporal accumulation of spectrum peaks," in *Proc. Interspeech 2010*, Sept. 2010, pp. 641–644.

[11] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. EUSIPCO 2011*, Aug. 2011, pp. 451–455.

[12] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2012*, Mar. 2012, pp. 409–412.

[13] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Magazine*, vol. 27, no. 6, pp. 66–80, 2010.

[14] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[15] F. Huang and T. Lee, "Robust pitch estimation using l_1-regularized maximum likelihood estimation," in *Proc. Interspeech 2012*, Sept. 2012.

[16] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, 1967.

[17] S. Ahmadi and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech, Audio Process.*, vol. 7, no. 3, pp. 333–338, May 1999.

[18] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis (Edited by W. B. Kleijn and K. K. Paliwal)*, pp. 495–518, 1995.

[19] F. Huang and T. Lee, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 99–109, Jan. 2013.

[20] F. Huang and T. Lee, "Sparsity-based confidence measure for pitch estimation in noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2012*, Mar. 2012, pp. 4601–4604.

[21] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[22] Y. T. Yeung, T. Lee, and C.-C. Leung, "Using dynamic conditional random field on single-microphone speech separation," accepted for *IEEE Int. Conf. Acoust., Speech, Signal Process. 2013*.

[23] L. R. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 399–418, Oct. 1976.

[24] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," 2000, Available online: http://www.speech.kth.se/wavesurfer/.

[25] M. Brookes, "VOICEBOX: Speech processing toolbox for MATLAB," Dec. 1997, Available online: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[26] V. Franc and V. Hlavac, "A novel algorithm for learning support vector machines with structured output spaces," Research Report K333–22/06, CTU–CMP–2006–04, May 2006, Program available online: http://cmp.felk.cvut.cz/~xfrancv/libqp/html/.

[27] F. Huang, T. Lee, and W. B. Kleijn, "A method of speech periodicity enhancement based on transform-domain signal decomposition," in *Proc. EUSIPCO 2010*, Aug. 2010, pp. 984–988.