SPARSITY-BASED CONFIDENCE MEASURE FOR PITCH ESTIMATION IN NOISY SPEECH

Feng Huang and Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR

ABSTRACT

In this paper, we propose a confidence measure for the pitch estimation method presented in a parallel paper [1]. The confidence measure is derived based on a sparse representation of the speech harmonic structure. The measurement quantitatively reflects the harmonicity associated with an estimated pitch. It is used to indicate the reliability of the results. Histogram is employed to illustrate the distribution of the measurement obtained from speech signals at low signal-to-noise ratios. It is shown that with the confidence measure, correct results can be effectively identified. By using the confidence measure, an adaptive algorithm for pitch estimation is proposed. Reliable pitch values obtained from preceding estimations are identified and used to predict the local pitch range for subsequent estimations. Parameter of the estimation algorithm is dynamically adjusted according to the predicted pitch range. Experimental results show that with the adaptive algorithm, pitch estimation accuracy is noticeably improved.

Index Terms— Robust pitch estimation, confidence measure, sparse representation, pitch range prediction

1. INTRODUCTION

Automatic determination of the fundamental frequency, i.e., pitch or F0, in noisy speech is an important basic problem for speech enhancement, robust speech recognition and many other areas of speech research. The most commonly used approach towards robust pitch estimation is to use complementary pitch cues [2, 3, 4]. However, the estimation accuracy is generally considered unsatisfactory, particularly when signal-to-noise ratio (SNR) is low. Besides, a common shortcoming of most existing methods is the absence of confidence measures for evaluating the reliability of the estimation results.

An effective confidence measure can be largely beneficial. The measurement can be used to identify estimation errors and thus enable subsequent process, e.g., speech enhancement, to prevent the errors from degrading system performance. With the confidence measure, pitch estimation accuracy can be improved. For instance, isolated pitch errors can be spotted and corrected. Moreover, the measurement can be used to indicate voicing status and hence can contribute to robust voiced/unvoiced decision (VUD) and voice activity detection (VAD).

In a parallel paper [1], we present a novel pitch estimation method for improving pitch estimation accuracy at low SNRs. Sparsity-related estimation approach [5, 6] is employed in the method. In this extension study, we propose a confidence measure for evaluating the reliability of the estimation results.

In the following, we will try to motivate the confidence measure by a brief review of the pitch estimation method. Pitch is estimated based on a temporal-spectral representation of the speech harmonic structure, namely *temporally accumulated peak spectrum* (TAPS). TAPS of the kth frame, $\mathbf{y}^{(k)}$, is computed by

$$\mathbf{y}^{(k)} = \mathbf{p}^{\left(k-\lfloor \frac{K}{2} \rfloor\right)} + \cdots + \mathbf{p}^{(k)} + \cdots + \mathbf{p}^{\left(k-\lfloor \frac{K}{2} \rfloor + K - 1\right)}, \quad (1)$$

where $\mathbf{p}^{(k)} \in \mathcal{R}^{M\times 1}$ is the peak spectrum *vector* obtained by retaining the peaks of the DFT magnitude spectrum and setting the other magnitudes to zero [7]. If $\mathbf{p}^{(k)}$ covers the full spectrum, then M equals the number of frequency bins. In Eq. (1), $\lfloor\cdot\rfloor$ is the floor function, "+" is entry-wise addition and K is the number of accumulated frames. Since pitch usually changes slowly in neighboring frames, in $\mathbf{y}^{(k)}$, harmonic-related peaks are concentrated around the fundamental frequency and its multiples. On the other hand, noise peaks in $\mathbf{y}^{(k)}$ are irregularly located and relatively small. A degree of robustness against noise can be gained [7].

In addition, prior speech information is utilized. Prior knowledge is incorporated as a large set of peak spectrum exemplars obtained from clean voiced speech. The exemplars over-completely represent all possible pitch values. An information matrix $\mathbf{A} = [\bar{\mathbf{p}}_1\ \bar{\mathbf{p}}_2\ \cdots\ \bar{\mathbf{p}}_N]$, where $\mathbf{A} \in \mathcal{R}^{M\times N}$ and $N \gg M$, is composed from the exemplars. Each column of $\mathbf{A}$ represents a peak spectrum exemplar. An accumulated peak spectrum $\mathbf{y}$ is then assumed to be represented as a sparse linear combination of the exemplars, i.e.,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}, \quad (2)$$

where $\mathbf{x} \in \mathcal{R}^{N\times 1}$ is a sparse weight vector with at most K non-zero elements. $\mathbf{v}$ represents the noise effect in the peak spectrum domain. With Gaussian assumption on the probability distribution of $\mathbf{v}$, $\mathbf{x}$ can be effectively estimated in a maximum likelihood sense [8]. Based on $\mathbf{x}$, the pitch is determined. The algorithm for obtaining $\mathbf{x}$ is provided in Appendix 1 for the convenience of the reader.

2. CONFIDENCE MEASURE FOR PITCH ESTIMATION

2.1. Voting weights

Let $\hat{\mathbf{x}} = [\hat{x}_1\hat{x}_2\cdots\hat{x}_n\ \cdots\hat{x}_N]^T$ denote the estimated weight vector. Each non-zero element in $\hat{\mathbf{x}}$ identifies a constituent exemplar for $\mathbf{y}$. Since a peak spectrum exemplar $\bar{\mathbf{p}}_n$ indicates a pitch value $f_0(n)$, all the constituent exemplars suggest a set of pitch candidates $\{{f_0}^{\mathrm{c}}(1)\cdots{f_0}^{\mathrm{c}}(n_f)\cdots{f_0}^{\mathrm{c}}(N_f)\}$.

There may be multiple constituent exemplars corresponding to a same pitch candidate. In practice, we also merge close pitch values via averaging to form a single candidate. For the candidate ${f_0}^{\mathrm{c}}(n_f)$, a voting weight $\hat{x}^{\mathrm{c}}_{n_f}$ is proposed as the sum of all associated non-zero elements in $\hat{\mathbf{x}}$, i.e.,

$$\hat{x}^{\mathrm{c}}_{n_f} = \sum_{\substack{1\le n\le N \\ \hat{x}_n>0 \\ f_0(n)=f_0^{\mathrm{c}}(n_f)}} \hat{x}_n. \tag{3}$$

The voting weights play an important role of identifying the major constituent exemplar(s). The fundamental frequency is determined from the dominant one(s). The candidate with the largest voting weight is selected as the estimated pitch $\hat{f}_0$. Let

$$n_f^* = \underset{n_f}{\operatorname{argmax}}\, \hat{x}^{\mathrm{c}}_{n_f}, \tag{4}$$

then the pitch is estimated as

$$\hat{f}_0 = f_0^{\mathrm{c}}(n_f^*). \tag{5}$$

Experimental comparison of the above method with conventional methods of robust pitch estimation is provided in [1]. It is shown that the above method can attend high estimation accuracy, especially at low SNRs.

2.2. The confidence measure

Numerical value of the largest voting weight $\hat{x}^{\mathrm{c}}_{n_f^*}$ provides an useful clue for evaluating the estimation result. A further insight can be gained from the sparse representation as in Eq. (2). In $\mathbf{y}$ of voiced speech, harmonic-related peaks are concentrated around the fundamental frequency and its multiples. In terms of the sparse representation, most of the non-zero elements in $\hat{\mathbf{x}}$ are expected to denote exemplars with the same or close pitch values. Therefore, if the sparse weight is correctly estimated, most of the non-zero elements in $\hat{\mathbf{x}}$ should contribute to $\hat{x}^{\mathrm{c}}_{n_f^*}$ and $\hat{x}^{\mathrm{c}}_{n_f^*}$ should be large. For incorrect results, it is observed that the sparse weight vector usually denotes exemplars with diverse pitch values. As a result, fewer non-zero elements in $\hat{\mathbf{x}}$ contribute to $\hat{x}^{\mathrm{c}}_{n_f^*}$ and $\hat{x}^{\mathrm{c}}_{n_f^*}$ is small. Similar situation is also observed with the unvoiced speech. Usually, when voicing status of the noisy speech is unknown, "pitch values" are also computed for the unvoiced frames. It will be beneficial if these "meaningless" results can be recognized. This is achievable by using $\hat{x}^{\mathrm{c}}_{n_f^*}$ as an indication. For unvoiced speech, there is no harmonic peak in $\mathbf{y}$. With the same estimation, the obtained $\hat{\mathbf{x}}$ normally denotes exemplars with different pitch values, and the corresponding $\hat{x}^{\mathrm{c}}_{n_f^*}$ turns out to be relatively small as well.

The value of $\hat{x}^{\mathrm{c}}_{n_f^*}$ shows a high correlation to the correctness or reliability of the estimated pitch. We use $\hat{x}^{\mathrm{c}}_{n_f^*}$ as an index for evaluating the result. Based on $\hat{x}^{\mathrm{c}}_{n_f^*}$, a confidence measure is proposed. It is computed by

$$P_{F0} = \frac{\hat{x}^{\mathrm{c}}_{n_f^*}}{\sum_{n_f} \hat{x}^{\mathrm{c}}_{n_f}} = \frac{\hat{x}^{\mathrm{c}}_{n_f^*}}{\sum_{n} \hat{x}_{n}}. \tag{6}$$

In Eq. (6), the normalization ensures that $0 < P_{F0} \le 1$. If P_{F0} is large, then $\hat{f}_0$ is likely to be a correct result and the estimation is reliable. When P_{F0} equals 1, all the non-zero elements in $\hat{\mathbf{x}}$ vote for the same pitch value. We treat that as the most confident estimation. If P_{F0} is small, then $\hat{f}_0$ may be incorrect or the signal may be unvoiced.

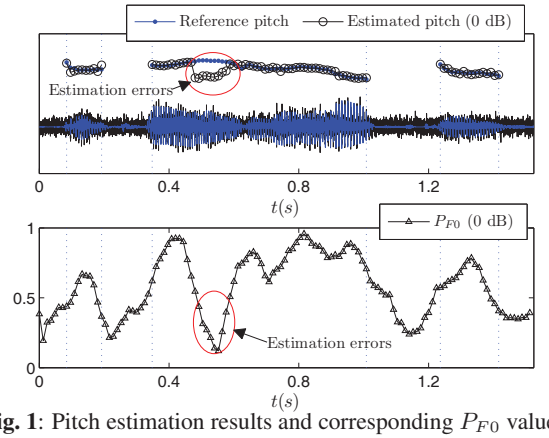


Fig. 1: Pitch estimation results and corresponding P_{F0} values

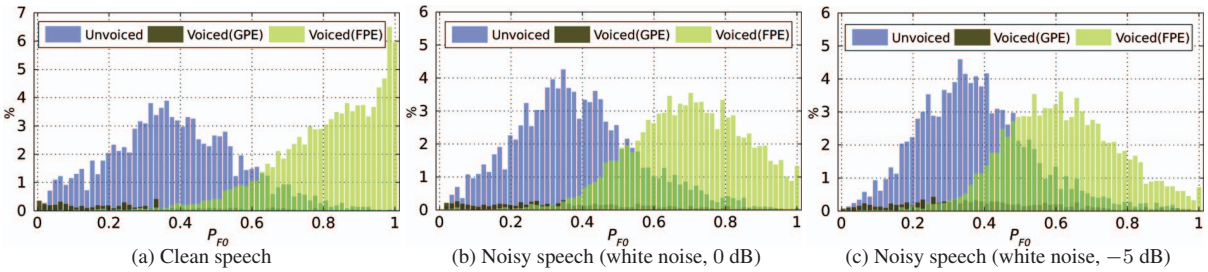
Fig. 1 shows an example of P_{F0} computed from real speech data. The upper figure shows, in an overlap manner, a clean speech segment (blue) and its counterpart (back) corrupted by 0 dB white noise. The reference pitch track and the pitch track estimated from the noisy signal are also illustrated. The lower figure shows the curve of P_{F0} obtained during the process of pitch estimation in the noisy signal. It can be seen that for the correct results, P_{F0} is large and mostly close to 1. For the errors, P_{F0} is small. It is also shown that for the unvoiced, P_{F0} is small as well.

2.3. Histogram analysis

We further conduct a statistical analysis on P_{F0}. We use histogram to illustrate the distributions of the numerical values of P_{F0} for voiced frames and unvoiced frames at different SNRs. For voiced frames, the values of P_{F0} for two types of results are also investigated in accordance with two performance metrics for pitch estimation: gross pitch error (GPE) and fine pitch error (FPE) [9]. For FPE, estimated pitch is within a close neighborhood ($\pm$16 Hz) of the true ones and it is treated as correct estimation, while GPE is on the contrary.

For the analysis, 80 utterances from 5 male and 5 female speakers are used. They are a subset of the CSLU-VOICES corpus [10]. The utterances were down-sampled to 8 kHz. Half of the utterances are used to train the parameters for the estimation algorithm, i.e., $\mathbf{A}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for (7) [1]. The other 40 are used for the following analysis. White noise is generated by software and added to the utterances. For each signal frame, $\mathbf{y}$ is computed and the sparse weight $\hat{\mathbf{x}}$ is estimated using (7). Based on $\hat{\mathbf{x}}$, the pitch is estimated and the corresponding P_{F0} is computed as described above. For voiced frames, the estimated pitch values are compared with the true ones. Accordingly, the P_{F0} results are divided into two groups, i.e., the FPE group and the GEP group.

For a group of P_{F0} results, a histogram is generated by counting the occurrences of the specific P_{F0} values. Fig. 2 shows the histograms of the FPE, GPE and unvoiced groups at



(a) Clean speech (b) Noisy speech (white noise, 0 dB) (c) Noisy speech (white noise, -5 dB)

Fig. 2: Histograms for the distributions of P_{F0}. (Histograms of the unvoiced group are normalized by the total number of unvoiced frames. Histograms of the GPE and FPE groups are normalized by the total number of voiced frames.)

Table 1: Optimal decision boundaries and corresponding accuracies

Task	Clean		0 dB		-5 dB	
	BND	ACC	BND	ACC	BND	ACC
FPE vs (GPE+U)	0.62	91.5%	0.52	85.9%	0.50	83.1%
FPE vs GPE	0.35	97.7%	0.37	94.5%	0.31	91.3%

different SNRs. The histograms obtained from clean speech are given in Fig. 2a. It can be seen that P_{F0} values of the FPEs are large and the majority is close to 1. For the GPEs and the unvoiced frames, most of the P_{F0} values are small. Fig. 2b and 2c show the histograms obtained from noisy speech. It can be seen that P_{F0} values of the unvoiced frames tend to follow a same distribution. For the FPEs, although the P_{F0} values generally become smaller when SNR decreases, the majority is still large and discriminable from the other groups. Following, we quantitatively analyze the discrimination level.

Consider two decision tasks: (1) Given P_{F0} of a frame, decide the corresponding estimated pitch is FPE or not (**FPE vs (GPE+U)**); (2) For a voiced frame, given P_{F0}, decide the corresponding estimated pitch is FPE or GPE (**FPE vs GPE**). From the above histograms, we can obtain optimal P_{F0} boundaries, with which the decision accuracies (on the analyzed data) for the respective tasks are the highest. The highest accuracies quantitatively reflects the discrimination levels of the respective groups. Table 1 gives the optimal boundaries (BND) and the corresponding accuracies (ACC). It can be seen that the discrimination levels are consistently high, even at low SNRs. It infers that based on P_{F0}, the correct results (FPE) can indeed be effectively identified.

3. ADAPTIVE PITCH ESTIMATION

3.1. Parameter adaptation

We demonstrate usefulness of the confidence measure by an application. Here, P_{F0} is used to identify reliable estimation results for further reduction of estimation errors. Specifically, an estimated pitch is recognized as reliable if the corresponding P_{F0} is larger than a threshold. Recent reliable pitch values are used to predict the local pitch range for subsequent estimations. With the predicted pitch range, the prior information matrix $\mathbf{A}$ is dynamically adjusted to cover the local pitch range. The adjusted information matrix is then used in the subsequent estimation. The adaptation of $\mathbf{A}$ is potentially beneficial. This is because originally $\mathbf{A}$ is designed to represent all possible pitch values [1]. However, for a local speech segment, the pitch is usually within a relative narrow range. A compact $\mathbf{A}$ is helpful for improving estimation accuracy, e.g., pitch errors such as doubling and halving can be avoided.

Table 2: Pitch estimation algorithm with adaptive parameter

Require:
- $\mathbf{A}$ The complete prior information matrix;
- N_{C} Number of cached results;
- $N_{\mathrm{R}}^{\min}$ Minimum number of reliable results for predicting the local pitch range, $N_{\mathrm{R}}^{\min} \leq N_{\mathrm{C}}$;
- P_{F0}^{th} P_{F0} threshold for identifying the reliable results.

Initial a queue Q of length N_{C}, $\mathsf{Q} \leftarrow 0, 0, \cdots, 0$
for each frame in the noisy speech **do**
  if Q.NumberOfNonZeroElements() $\geq N_{\mathrm{R}}^{\min}$ **then**
    $\mathrm{F0}_{\min} \leftarrow \min(\mathsf{Q}.\mathrm{NonZeroElements}()) - C$
    $\mathrm{F0}_{\max} \leftarrow \max(\mathsf{Q}.\mathrm{NonZeroElements}()) + C$
    $\mathbf{A}_{\mathrm{loc}} \leftarrow$ From $\mathbf{A}$ select the exemplars whose F0 within $[\mathrm{F0}_{\min}\ \mathrm{F0}_{\max}]$
  else
    $\mathbf{A}_{\mathrm{loc}} \leftarrow \mathbf{A}$
  end if
  Compute $\mathbf{y}$ using Eq. (1)
  Estimate the sparse weigh $\mathbf{x}$ with $\mathbf{A}_{\mathrm{loc}}$ using (7)
  Obtain $\hat{f}_0$ using Eq. (3), (4) and (5)
  Compute P_{F0} using Eq. (6).
  if $P_{F0} \geq P_{F0}^{\mathrm{th}}$ **then**
    $\mathsf{Q}.\mathrm{Enqueue}(\hat{f}_0)$
  else
    $\mathsf{Q}.\mathrm{Enqueue}(0)$
  end if
end for

Table 2 shows the adaptive algorithm. In the algorithm, local pitch range is predicted as the range between the minimum and maximum (extended by C Hz) of the recent reliable results. If there is not enough reliable results, the algorithm falls back to the original one, i.e., the complete $\mathbf{A}$ is used.

3.2. Evaluation

Performance of the adaptive algorithm (**Adpt**) is evaluated and compared with the original algorithm (**Orig**) where $\mathbf{A}$ is complete and fixed. A previously proposed approach [7], where pitch is estimated from the autocorrelation of $\mathbf{y}$ (**AutoC**), is also involved in the comparison. The 40 utterances for the above histogram analysis are also used for this evaluation. Three types of noise, i.e., white noise, non-stationary (NS) white noise and car noise, are used. NS-white noise was obtained by randomly changing the variance of white noise every 8 ms in the range of $[\sigma^2, 5\sigma^2]$ with $\sigma^2 = 1$.

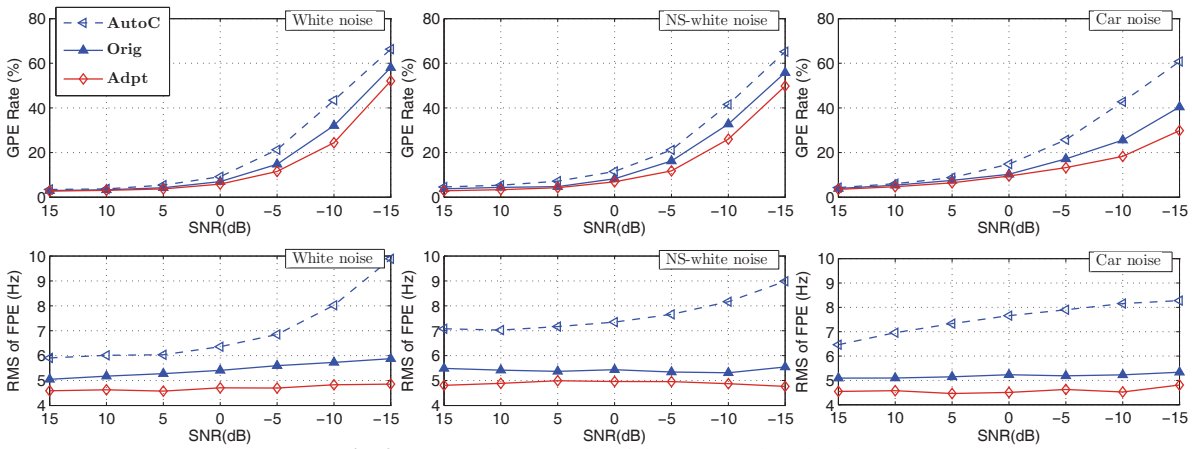


Fig. 3: GPE and FPE results of the compared methods.

Car noise (VOLVO-340, 120 km/h) was obtained from the NOISEX-92 data set. The parameters, $\mathbf{A}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, trained for the histogram analysis are also used. The number of exemplars in $\mathbf{A}$ is $N = 1024$. The dimension of peak spectrum vector is $M = 102$, covering the frequency range of 0 Hz to 800 Hz. For the **Adpt** algorithm as in Table 2, we empirically set $N_{\mathrm{C}} = 10$, $N_{\mathrm{R}}^{\min} = 5$, $P_{F0}^{\mathrm{th}} = 0.70$ and $C = 30$ Hz.

The speech signals are degraded with noise at various SNRs. The pitch is estimated and compared with the reference pitch, which is obtained via manually labeling the clean waveforms. For GPE, the error rate in percentage is calculated. For FPE, the root mean square (RMS) of the estimation deviation is computed [1, 9]. Fig. 3 shows the GPE and FPE results. It can be seen that with the **Adpt** algorithm, estimation accuracy is noticeably improved, specially at low SNRs. GPE rates of the **Adpt** algorithm are the lowest for all SNR conditions. Moreover, the average FPE result of the **Adpt** algorithm is 0.62 Hz lower than that of the **Orig** algorithm. The improvement confirms the effectiveness of using preceding reliable results and infers that the reliable pitch results are indeed effectively identified.

4. CONCLUSIONS

A confidence measure has been proposed to evaluate the estimation results of a new pitch estimation method. The measurements effectively reflect the reliability of the estimation results. Histogram analysis confirms that with the confidence measure, correct results (FPE) can be effectively discriminated from gross errors and those estimated from unvoiced speech. By using the confidence measure, an adaptive algorithm for pitch estimation was established. With the adaptive method, pitch estimation accuracy was noticeably improved. The improvement confirms usefulness of the confidence measure. As for further application, the confidence measurement can be used in existing VUD and VAD algorithms to improve detection accuracy at low SNRs.

ACKNOWLEDGMENT

This research is partially supported by the General Research Funds (Ref: CUHK 413507 & CUHK 414108) from the Hong Kong Research Grants Council, and a project grant from the Shun Hing Institute of Advanced Engineering, CUHK.

Appendix 1. SPARSE WEIGHT ESTIMATION

Given the probability density function of $\mathbf{v}$ as $\varphi(\mathbf{v}) = \mathcal{N}(\mathbf{v}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the maximum likelihood (ML) estimation of $\mathbf{x}$ for Eq. (2) can be obtained by minimizing the negative log-likelihood function $-\log \varphi(\mathbf{y} - \mathbf{A}\mathbf{x})$. Consider the sparsity constraint on $\mathbf{x}$, an l_1 regularization term is imposed [1]. The l_1-regularized ML estimation of $\mathbf{x}$ is obtained by

$$\begin{aligned} \min_{\mathbf{x}} \quad & (\mathbf{A}\mathbf{x} - \mathbf{y} + \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{A}\mathbf{x} - \mathbf{y} + \boldsymbol{\mu}) \\ \text{subject to} \quad & \|\mathbf{x}\|_1 \leq K \text{ and } \mathbf{x} > \mathbf{0}. \end{aligned} \tag{7}$$

5. REFERENCES

[1] F. Huang and T. Lee, "Robust pitch estimation using l_1-regularized maximum likelihood estimation," submitted to *ICASSP 2012*.

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

[3] M. Heckmann and et al., "Combining rate and place information for robust pitch extraction," in *Proc. Interspeech '07*, Aug 2007, pp. 2765–2768.

[4] L. N. Tan and A. Alwan, "Noise-robust f0 estimation using snr-weighted summary correlograms from multi-band comb filters," in *Proc. ICASSP '11*, May 2011, pp. 4464 –4467.

[5] J. F. Gemmeke and B. Cranen, "Noise reduction through compressed sensing," in *Proc. Interspeech '08*, Sept 2008, pp. 1785–1788.

[6] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," in *Proc. ICASSP '09*, Apr 2009, pp. 4125–4128.

[7] F. Huang and T. Lee, "Pitch estimation in noisy speech based on temporal accumulation of spectrum peaks," in *Proc. Interspeech '10*, Sept 2010, pp. 641–644.

[8] A. Zymnis, S. Boyd, and E. Candes, "Compressed sensing with quantized measurements," *IEEE SPL*, vol. 17, no. 2, pp. 149 –152, Feb 2010.

[9] L. R. Rabiner and et al., "A comparative performance study of several pitch detection algorithms," *IEEE Trans. ASSP*, vol. 24, pp. 399–418, Oct 1976.

[10] A. Kain, "CSLU: VOICES," *Linguistic Data Consortium, Philadelphia*, 2006.