ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Robust Pitch Estimation Using l_1-regularized Maximum Likelihood Estimation

Feng Huang and Tan Lee
Department of Electronic Engineering, The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR
{fhuang,tanlee}@ee.cuhk.edu.hk

10.21437/Interspeech.2012-137

Abstract

This paper presents a new method of robust pitch estimation using sparsity-based estimation techniques. The method is developed based on sparse representation of a temporal-spectral pitch feature. The robust pitch feature is obtained by accumulating spectral peaks over consecutive frames. It is expressed as a sparse linear combination of an over-complete set of peak spectrum exemplars. The probability distribution of the noise is assumed to be Gaussian with non-zero mean. The weights of the linear combination are estimated by maximizing the likelihood of the feature under sparsity constraint. The sparsity constraint is incorporated as an l_1 regularization term. From the estimated weights, the major constituent exemplars are identified and the fundamental frequency is determined. Experimental results show that, with this method, pitch estimation accuracy is significantly improved, particularly at low signal-to-noise ratios.

Index Terms: Robust pitch estimation, speech sparsity, l_1 regularization, peak spectrum

1. INTRODUCTION

Pitch, measured in terms of fundamental frequency (F0), is one of the most important attributes of human speech. Important high-level linguistic information, for examples, intonation, lexical tones, stress and focus, is conveyed by the pitch contour of a spoken utterance. Pitch is particularly essential for tonal languages, where lexical or grammatical meanings of words are distinguished by pitch contour shape and/or pitch range.

Detecting the fundamental frequency of a speech signal is an important basic problem in speech enhancement, robust speech/speaker recognition and many other areas of speech research. With the presence of noise, both time-domain periodicity and frequency-domain harmonicity of speech signal are distorted. Conventional methods [1] become less effective and inaccurate. A commonly used approach towards robust pitch estimation is to use complementary pitch cues. For instance, ACF and AMDF functions were jointly used in [2], while rate and place pitch cues were jointly used in [3]. In [4], multi-band correlograms were used. However, the estimation performance is generally considered unsatisfactory, particularly when signal-to-noise ratio (SNR) is low.

One important approach to improving noise robustness is to utilize prior information. The methods that use hidden Markov models for multiple-pitch tracking [5, 6] are examples of such approach. However, there have been very few studies on effective exploitation of prior speech information for single pitch estimation.

In this paper, we present a novel method of single pitch estimation with the goal of improving estimation accuracy at low SNRs. The method effectively incorporates prior knowledge about speech harmonic structure and utilizes sparsity of the structure in pitch estimation. Peak spectrum [7] is used to describe the harmonic structure. A peak spectrum is obtained by retaining the peaks of the DFT magnitude spectrum and setting the other magnitudes to zero. Representative peak spectra are learned from clean speech and recruited to form a large set of exemplars. For noisy speech, a temporal-spectral pitch feature is employed. The feature vector is obtained by accumulated the peak spectra of a few neighboring frames. It is then expressed as a sparse linear combination of the exemplars. With the assumption that probability distribution of the noise effect in the peak spectrum domain is Gaussian, the weights for the sparse representation are estimated via l_1-regularized maximum likelihood (ML) estimation. From the estimated weights, the constituent exemplars are identified and the fundamental frequency is determined. Our experimental results confirm that pitch estimation accuracy can be significantly improved with this new method.

In Section 2, we first briefly review the temporal-spectral pitch feature and describe the sparse representation. In Section 3, we present the l_1-regularized ML estimation method for pitch estimation. In Section 4 and 5, implementation issues are discussed and experimental results are provided. Finally in Section 6, we summarize and conclude this study.

2. SPARSE REPRESENTATION OF SPEECH HARMONIC STRUCTURE

2.1. A temporal-spectral representation

In our previous study [7], we demonstrated that the *peak spectrum* was effective and robust in revealing inter-frame similarity of the speech harmonic structure. Let $\mathbf{s}^{(k)} = [S^{(k)}(1)\ S^{(k)}(2)\ \cdots S^{(k)}(m)\ \cdots\ S^{(k)}(M)]^T$ be the magnitude spectrum vector of the kth frame. M is the number of frequency bins. The corresponding peak spectrum vector $\mathbf{p}^{(k)} = [P^{(k)}(1)\ P^{(k)}(2)\cdots P^{(k)}(m)\cdots P^{(k)}(M)]^T$ is derived from $\mathbf{s}^{(k)}$ by

$$P^{(k)}(m) = \begin{cases} S^{(k)}(m) & \text{if } \Delta S^{(k)}(m) < 0 \text{ and } \nabla S^{(k)}(m) > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

where m is frequency-bin index, Δ and ∇ denote the forward and backward difference operators, respectively.

Based on the peak spectrum, a robust temporal-spectral

representation of the speech harmonic structure, namely *temporally accumulated peak spectrum* (TAPS), was derived [7]. TAPS of the kth frame, denoted as $\mathbf{y}^{(k)}$, is obtained by accumulating peak spectra of a few neighboring frames, i.e.,

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

where "+" is the entry-wise addition, $\lfloor \cdot \rfloor$ is the floor function, and K is the number of accumulated frames. In speech signals, the fundamental frequency usually does not change rapidly in connected frames. Thus in $\mathbf{y}^{(k)}$, harmonic-related peaks are concentrated around the fundamental frequency and its multiples, while noise peaks are irregularly located and relatively small. The autocorrelation of $\mathbf{y}$ could be used for pitch estimation in noisy speech [7].

2.2. Sparse representation of TAPS with prior information

As for human speech, the fundamental frequency covers a wide range. The peak spectra exhibit various structures regarding the frequency spacing between the harmonic peaks. However, for speech signals with the same fundamental frequency, the peak spectrum structures are considered as similar since the frequency spacing between harmonic peaks is the same. Given a set of peak spectrum exemplars which over-completely represent all possible pitch values, an observed peak spectrum can be expressed by one or a couple of the exemplars, e.g., linear interpolation of the specific exemplars to obtain desired magnitudes. The accumulated peak spectrum $\mathbf{y}$ can then be sparsely composed from a small number of the representative exemplars. (For $\mathbf{y}^{(k)}$, the superscript $^{(k)}$ is hereafter omitted for simplicity.)

Let $\mathbf{A} = [\bar{\mathbf{p}}_1\ \bar{\mathbf{p}}_2\ \cdots\ \bar{\mathbf{p}}_n \cdots\ \bar{\mathbf{p}}_N]$, where $\mathbf{A} \in \mathcal{R}^{M\times N}$ and $N \gg M$, be a prior information matrix, with each column being a peak spectrum exemplar, $\mathbf{y}$ is represented as a sparse linear combination of the exemplars, i.e.,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}, \quad (3)$$

where $\mathbf{x} \in \mathcal{R}^{N\times 1}$ is a sparse weight vector, $\mathbf{v} \in \mathcal{R}^{M\times 1}$ represents the noise effect in the peak spectrum domain. With an over-complete $\mathbf{A}$, we assume that the number of constituent exemplars is at most K according to the definition in Eq (2).

3. ROBUST PITCH ESTIMATION

3.1. Sparse weight estimation under Gaussian assumption

For pitch estimation, the goal is to obtain $\mathbf{x}$ given an observed $\mathbf{y}$. With the sparse weight $\mathbf{x}$ estimated, the peak spectrum components for $\mathbf{y}$ can be identified and the fundamental frequency can be determined.

We assume that the probability distribution of $\mathbf{v}$ is Gaussian. Since the peak spectrum is derived from the magnitude spectrum, some dimensions of $\mathbf{v}$, e.g., the elements whose corresponding frequency bins are in-between speech harmonics, may always tend to be non-negative. Therefore, the mean of the Gaussian distribution should not be zero. It is assumed to be non-zero. Denote the density function of $\mathbf{v}$ as

$$\varphi(\mathbf{v}) = \mathcal{N}(\mathbf{v}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (4)$$

where $\boldsymbol{\mu} \neq \mathbf{0}$ and $\boldsymbol{\Sigma} \succ \mathbf{0}$ are the mean vector and covariance matrix, respectively. With $\varphi(\mathbf{v})$, the sparse weight vector $\mathbf{x}$ is estimated by maximizing the likelihood of $\mathbf{y}$. Define the conditional probability of $\mathbf{y}$ given $\mathbf{x}$ as [8]

$$p(\mathbf{y}|\mathbf{x}) = \varphi(\mathbf{y} - \mathbf{A}\mathbf{x}). \quad (5)$$

Correspondingly, the negative log-likelihood function is

$$f(\mathbf{y}|\mathbf{x}) = -\log[\varphi(\mathbf{y} - \mathbf{A}\mathbf{x})]. \quad (6)$$

$f(\mathbf{y}|\mathbf{x})$ is used as the objective function for the ML estimation of $\mathbf{x}$. Since $\mathbf{y}$ is known, we denote $f(\mathbf{y}|\mathbf{x})$ as $f_{\mathrm{ml}}(\mathbf{A}\mathbf{x})$. With Eq. (4), it can be derived that

$$f_{\mathrm{ml}}(\mathbf{A}\mathbf{x}) = c \cdot (\mathbf{A}\mathbf{x} - \mathbf{y} + \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{A}\mathbf{x} - \mathbf{y} + \boldsymbol{\mu}), \quad (7)$$

where $c = \log\left((2\pi)^M |\boldsymbol{\Sigma}|\right)$. The ML estimation of $\mathbf{x}$ is thus obtained by

$$\min_{\mathbf{x}} f_{\mathrm{ml}}(\mathbf{A}\mathbf{x}). \quad (8)$$

The weight vector obtained from (8) is usually not sparse, i.e., most of the elements are non-zero. An l_1 regularization term $\lambda\|\mathbf{x}\|_1$ is added to the objective function. The l_1-regularized ML estimation is formulated as

$$\min_{\mathbf{x}}\ f_{\mathrm{ml}}(\mathbf{A}\mathbf{x}) + \lambda\|\mathbf{x}\|_1, \quad (9)$$

where the parameter $\lambda > 0$ controls the degree of sparsity in $\mathbf{x}$. In Eq. (7), $f_{\mathrm{ml}}(\mathbf{A}\mathbf{x})$ is a quadratic function of $\mathbf{x}$, and it is convex given $c > 0$. Therefore, (9) can be effectively solved [9]. Moreover, with appropriate parameter correspondence, the above unconstrained optimization problem can be converted into a constrained one [10]. In this study, we use the following formulation to estimate $\mathbf{x}$,

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f_{\mathrm{ml}}(\mathbf{A}\mathbf{x}) \\ \text{subject to} & \|\mathbf{x}\|_1 \leq K \text{ and } \mathbf{x} > \mathbf{0}. \end{array} \quad (10)$$

The constraint $\|\mathbf{x}\|_1 \leq K$ is set according to the assumption that there are at most K constituent exemplars for $\mathbf{y}$.

In compressed sensing [8] as well as some recent studies [11, 12] in the speech area, similar approaches were used with the assumption that $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$. In our application, $\boldsymbol{\mu}$ is non-zero.

3.2. Pitch estimation from the sparse weight vector

The estimated weight $\hat{\mathbf{x}} = [\hat{x}_1\ \hat{x}_2\ \cdots\ \hat{x}_n\ \cdots\ \hat{x}_N]^T$ has a limited number of non-zero elements, each of which corresponds to a constituent peak spectrum exemplar. Since an exemplar $\bar{\mathbf{p}}_n$ indicates a pitch value $f_0(n)$, by scanning through all the non-zero weights, a set of pitch candidates $\{{f_0}^{\mathrm{c}}(1) \cdots {f_0}^{\mathrm{c}}(n_f) \cdots {f_0}^{\mathrm{c}}(N_f)\}$ is obtained. Note that there may be multiple exemplars corresponding to a same candidate pitch value. In practice, we also merge close pitch values to generate a single candidate. The fundamental frequency is determined from the dominant exemplar(s). A voting weight $\hat{x}^{\mathrm{c}}_{n_f}$ for the candidate ${f_0}^{\mathrm{c}}(n_f)$ is computed by summing up all associated non-zero elements in $\hat{\mathbf{x}}$, i.e.,

$$\hat{x}^{\mathrm{c}}_{n_f} = \sum_{\substack{1\leq n\leq N \\ \hat{x}_n>0 \\ f_0(n)=f_0^{\mathrm{c}}(n_f)}} \hat{x}_n. \quad (11)$$

The candidate that has the largest voting weight is chosen as the estimated pitch $\hat{f}_0$, i.e.,

$$\hat{f}_0 = f_0^{\mathrm{c}}(n_f^*), \quad (12)$$

where

$$n_f^* = \underset{n_f}{\operatorname{argmax}}\, \hat{x}^{\mathrm{c}}_{n_f}. \quad (13)$$

4. IMPLEMENTATION ASPECTS

4.1. Prior information matrix A

The peak spectrum exemplars are learned from voiced segments of clean speech. They are required to cover all possible pitch values. A clustering procedure is carried out to obtain the representatives. Specifically, from the training utterances, a large number of peak spectrum vectors are computed and divided into groups. Each group corresponds to a legitimate pitch value. Then, a k-means clustering process is performed on each group. The resulted clustering centers are recruited as the members of the exemplar set. The number of clusters for each pitch group is pre-determined in accordance with the number of available observations.

4.2. Gaussian density of v

Parallel clean and noisy TAPS vectors are used to obtain the Gaussian distribution of $\mathbf{v}$. They are computed from clean training utterances and their noise-corrupted counterparts. The training observations of $\mathbf{v}$ are obtained as the difference between the noisy TAPS vectors and the corresponding clean ones. From the training observations of $\mathbf{v}$, the mean vector and covariance matrix are computed. Since some dimensions of the peak spectrum vector are zero, the estimated mean and variances in certain dimensions may be zero as well. To avoid ill conditioning of the covariance matrix, a floor value of $\frac{1}{2\pi}$ is imposed to the variances. The floor value also helps to ensure $c > 0$ in Eq. (7).

5. PERFORMANCE EVALUATION

The proposed method (**TAPS-l_1-ML**) is evaluated and compared with both conventional and recent robust pitch estimation methods, including the cepstrum method (**CEP**) [13], the time-domain weighted autocorrelation method (**WAutoC**) [2], the robust algorithm for pitch tracking (**RAPT**) [14, 15], the pitch estimation filter with amplitude compression (**PEFAC**) [16, 15]. **CEP** and **WAutoC** are considered as conventional robust methods. **PEFAC** is chosen as a representative of the most recent spectral-domain methods, while **RAPT** represents methods that involve single-pitch tracking. A previously proposed approach [7], where pitch is estimated directly from the autocorrelation of TAPS (**TAPS-AutoC**), is also involved in the comparison.

80 gender-balanced utterances from the CSLU-VOICES corpus [17] are used. The utterances were down-sampled to 8 KHz. Half of them are used to train the exemplar matrix and the Gaussian density. The other 40 utterances are used for evaluation. The frame size and frame shift are 24 ms and 12 ms. For spectral analysis, 1024-point FFT with Hamming window is used.

We test with three types of noise, i.e., white noise, non-stationary (NS) white noise and car noise. White noise was generated with MATLAB functions. NS-white noise was obtained by randomly changing the variance of the white noise every 8 ms in the range of $[\sigma^2, 5\sigma^2]$ with $\sigma^2 = 1$. Car noise (VOLVO-340, 120 km/h) was obtained from the NOISEX-92 database.

Table 1: GPE and FPE results on clean speech

	CEP		**WAutoC**		**PEFAC**	**RAPT**	**TAPS-AutoC**	**TAPS-l_1-ML**
	24ms	60ms	24ms	60ms				
GPE	9.1%	8.6%	6.3%	6.8%	8.2%	4.4%	2.9%	2.5%
FPE (Hz)	5.86	6.04	5.17	5.79	5.28	4.68	5.96	4.83

Pitch estimation accuracy is evaluated in terms of *gross pitch error* (GPE) and *fine pitch error* (FPE) [18]. A computed $\hat{f}_0$ value is regarded as a GPE if $\hat{f}_0 \notin \left[f_0^{\min} - 16\text{Hz},\ f_0^{\max} + 16\text{Hz}\right]$. Otherwise, it is regarded as an FPE. $f_0^{\min}$ and $f_0^{\max}$ are the minimum and maximum reference pitch values over the analyzed signal segment. For each test utterance at a specified SNR, 10 noise-corrupted versions are created and evaluated. For GPE, the error rate in percentage is calculated. For FPE, the root mean square (RMS) of the deviation $|\hat{f}_0 - (f_0^{\max} - f_0^{\min})/2\ |$ is computed. For result verification, voicing status and reference pitch were obtained via manually labeling the clean waveforms. The reference pitch tracks were also cross-validated with the concurrent laryngograph signals.

For TAPS-based algorithms, the number of accumulated frames K is set to 4. The dimension of peak spectrum vector is $M = 102$, which covers the frequency range from 0 Hz to 800 Hz. The total number of exemplars in $\mathbf{A}$ is $N = 1024$. With $K = 4$, each pitch value is estimated from a signal segment of 60 ms. The **CEP** and **WAutoC** methods are tested with a frame length of 24 ms as well as 60 ms for fair comparison. **PEFAC** and **RAPT** are tested with frame length of 60 ms and frame shift of 12 ms.

Table 1 gives the GPE and FEP results obtained with clean speech. Fig. 1 shows the results obtained with speech degraded by the three types of noise at various SNRs.

From the results, it can be seen that **TAPS-l_1-ML** consistently outperforms the conventional methods, especially at low SNRs. For white noise, GPE rate of **TAPS-l_1-ML** is also noticeably lower than the recent **PEFAC** method. For instance, with white noise at -10 dB, GPE rate of the **TAPS-l_1-ML** method is 27.6% lower than **WautoC**(60 ms) and 7.2% lower than **PEFAC**. For car noise, **PEFAC** and **TAPS-l_1-ML** outperform the other methods, especially at SNR $\leq$ 0 dB; In these low SNR conditions (car noise), **PEFAC** tends to perform slightly better than **TAPS-l_1-ML**. This is because **PEFAC** uses amplitude compression that can attenuate the narrow-band car noise [16]. From the FPE results, it can be observed that **TAPS-l_1-ML** has significantly lower FPE deviation at low SNRs ($<$ 0 dB). It is also noticed that the FPE results of **TAPS-l_1-ML** are relatively constant across all SNR levels. It indicates that the **TAPS-l_1-ML** method provides more accurate estimation of the pitch value.

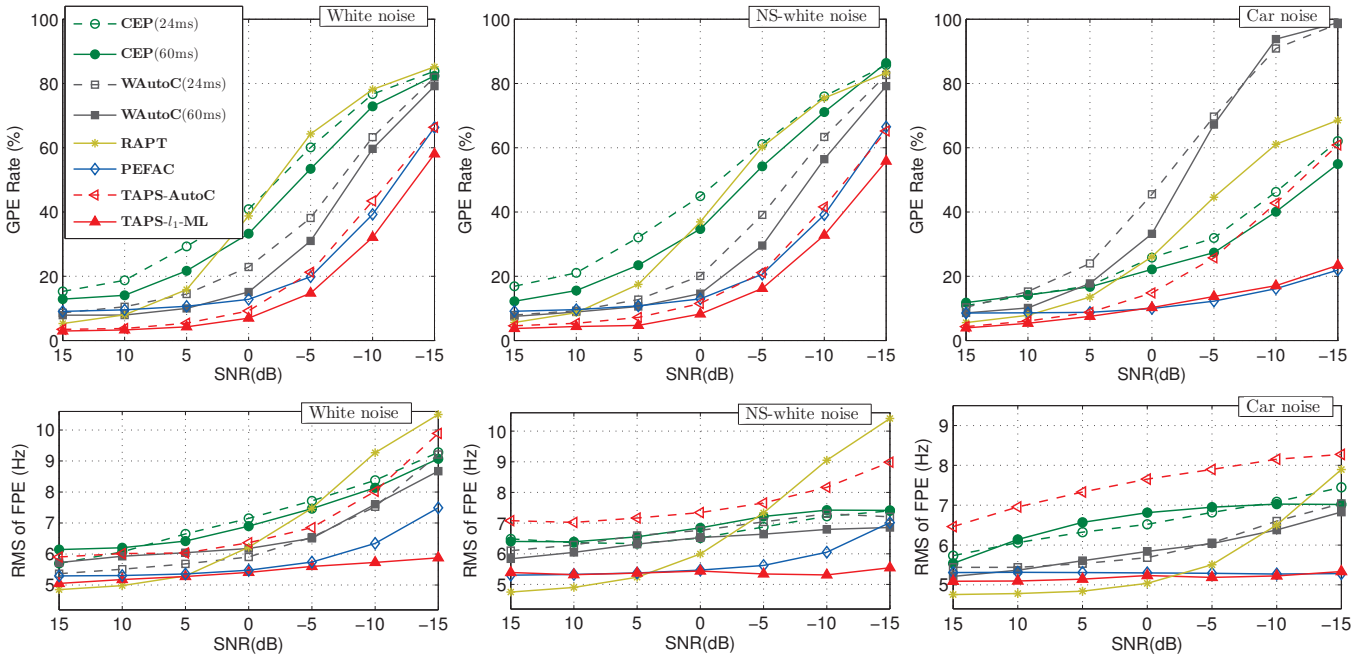

Figure 1: GPE and FPE results of the evaluated methods.

6. CONCLUSIONS

A new method of robust pitch estimation has been proposed. The method consistently outperforms the conventional methods at various SNRs and demonstrates noticeable advantages over the recently proposed spectral-domain method PEFAC. The main credit for the improvement attaches to the effective incorporation of prior information and the use of sparse representation and estimation techniques. In this study, the formulation of l_1-regularized ML estimation was derived with the assumption that probability distribution of the noise effect is non-zero mean Gaussian. Experimental results confirm that this assumption is applicable to practical situations. To model the probability distribution of the noise effect more accurately, Gaussian mixture model can be used.

ACKNOWLEDGMENT

This research is partially supported by the General Research Funds (Ref: CUHK 414010 & CUHK 413811) from the Hong Kong Research Grants Council, and a project grant from the Shun Hing Institute of Advanced Engineering, CUHK.

7. REFERENCES

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

[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] M.-Y. Wu, D.-L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. SAP*, vol. 11, no. 3, pp. 229–241, May 2003.

[6] Z.-Z. Jin and D.-L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. ASLP*, vol. 19, no. 5, pp. 1091–1102, Jul. 2011.

[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] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, Dec. 2010.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

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

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

[13] A. M. Noll, "Cepstrum pitch determination," *JASA*, vol. 41, pp. 293–309, 1967.

[14] 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.

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

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

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

[18] 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.