SPEAKER IDENTIFICATION USING SPARSELY EXCITED SPEECH SIGNALS AND COMPRESSED SENSING

Anthony Griffin, Eleni Karamichali and Athanasios Mouchtaris

Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH-ICS)
and Department of Computer Science, University of Crete,
Heraklion, Crete, Greece
{agriffin, karamih, mouchtar}@ics.forth.gr

ABSTRACT

Compressed sensing samples signals at a much lower rate than the Nyquist rate if they are sparse in some basis. Using compressed sensing theory to reconstruct speech signals was recently proposed, assuming speech signals are sparse in the excitation domain if they are modelled using the source/filter model. In this paper, the compressed sensing theory for sparsely excited speech signals is applied to the specific problem of speaker identification, and is found to provide encouraging results using a number of measurements as low as half of the signal samples. In this manner, compressed sensing theory allows the use of less samples to achieve accurate identification, which in turn would be beneficial in several sensor network related applications. Additionally, enforcing sparsity on the excitation signal is shown to provide identification accuracy which is more robust to noise than using the noisy signal samples.

1. INTRODUCTION

Speaker identification is the task of determining an unknown speaker's identity. In this paper, text-independent speaker identification is performed based solely on a speaker's voice. Speaker identification is achieved by performing a one-to-many match among the unknown voice signal and the previously available speech database of multiple speakers, assuming that the unknown speaker belongs in this dataset. The paper focuses on the possibility of performing speaker identification by applying the recently-proposed compressed sensing theory.

Compressed sensing (CS) [1–3] seeks to represent a signal using a small number of linear, non-adaptive measurements. Usually the number of measurements is much lower than the number of samples needed if the signal is sampled at the Nyquist rate. Thus, compressed sensing combines compression and sampling of a signal into one low-complexity step. An important restriction is that compressed sensing requires that the signal is sparse in some basis—in the sense that it is a linear combination of a small number of basis functions—in order to correctly reconstruct the original signal.

The reasons for examining the applicability of compressed sensing theory to the speaker identification problem are twofold. Firstly, compressed sensing theory achieves the reconstruction of a sparse signal using only a fraction of the number of samples dictated by the Nyquist theorem. Therefore, in a sensor-network scenario, the measurement operation could be performed locally and the few measurements in each time frame could be transmitted to a base station for further processing. From a different point of view, the second reason is due to the aforementioned sparsity restriction: by forcing the signal to be sparse in some basis, a noisy signal may be more robustly reconstructed. This is similar to signal denoising by low-rank modelling. In this case, the signal sparsity is an important factor, since the CS reconstruction will only be valid for signals which are initially sparse in some domain. Thus, in this second approach, we are interested in testing whether compressed sensing-based speaker identification results in a more robust identification than when directly using the signal's samples to perform the identification.

A key question is whether a speech signal can be considered to be sparse in some sense. For audio signals, we recently showed that their sinusoidally modelled component can be considered to be sparse, and compressed sensing theory was applied to low-bitrate audio coding [4]. For speech signals, compressed sensing was recently applied to a sparse representation using the source/filter model in [5] for speech coding, and encouraging preliminary results were obtained. In this paper we extend the work of [5] by applying the proposed methodology to the problem of text-independent speaker identification. In that work, it was found that applying compressed sensing theory to speech signals modelled using the source/filter model, and assuming a sparse excitation, resulted in accurate estimation of the filter part (spectral envelope) of the speech signal. For the filter part, a codebook for the speaker was used. Consequently, in this paper we create a filter codebook for each of the speakers in the database, and the identification process is based on selecting the speaker in the database corresponding to the codebook that results in the best compressed sensing reconstruction. It is shown that the percentage of correct identification using compressed sensing theory can reach 80% on average using a number of measurements which are as low as half of the signal's samples. When additive noise is used, the performance of compressed sensing-based identification is shown to be quite robust, with reference to a baseline GMM-based approach [6] for this task.

It is relevant at this point to mention the work in [7],

This work was funded in part by the PEOPLE-IAPP "AVID-MODE" grant within the 7th European Community Framework Program, and in part by the FORTH-ICS internal RTD program "AmI: Ambient Intelligence and Smart Environments".

where compressed sensing is used in the same context of the source/filter model, in order to derive a sparse residual for speech signals, when their filter part is known. In that case, the signal is sampled at the Nyquist rate (so as to derive the filter part), and the use of compressed sensing is to derive a sparse excitation signal as an alternative to multi-pulse excitation coders. In contrast, in this paper (as in [5]), the signal is in fact sampled at a rate which is significantly smaller compared to Nyquist, and consequently the codebook for the filter part is necessary in order to perform the CS-based reconstruction.

The proposed work mainly examines the applicability of CS-based reconstruction to speaker identification due to the advantage of sub-Nyquist sampling. At the same time, the fact that this method is shown here to be noise-robust, relates to previous work on feature extraction for robust speaker identification. A great amount of research efforts has focused on deriving robust features from the noisy speech signals, which can then be used for improving speaker identification performance as in [8–11]. Alternatively, the noise robust speaker identification problem can be examined as a problem of mismatched testing and training conditions as in [12, 13].

The remainder of the paper is organized as follows. In Section 2, the basics of compressed sensing are reviewed. In Section 3, the compressed sensing methodology for sparsely excited speech signals of [5] is summarized. Section 4 presents our approach on speaker identification using the compressed sensing reconstruction for sparsely excited speech signals. Identification results are given in Section 5, and concluding remarks are given in Section 6.

2. COMPRESSED SENSING

We sample the speech signal $x(t)$ at the Nyquist rate and process it in frames of N samples. Each frame is then an $N \times 1$ vector $\mathbf{x}$, which can be represented as

$$\mathbf{x} \; = \; \mathbf{\Psi X}, \tag{1}$$

where $\mathbf{\Psi}$ is an $N \times N$ matrix whose columns are the similarly sampled basis functions $\Psi_i(t)$, and $\mathbf{X}$ is the vector that chooses the linear combinations of the basis functions. $\mathbf{X}$ can be thought of as $\mathbf{x}$ in the domain of $\mathbf{\Psi}$, and it is $\mathbf{X}$ that is required to be sparse for compressed sensing to perform well. We say that $\mathbf{X}$ is K-sparse if it contains only K non-zero elements. In other words, $\mathbf{x}$ can be exactly represented by the linear combination of K basis functions.

It is also important to note that compressed sensing will also recover signals that are not truly sparse, as long these signals are highly *compressible*, meaning that most of the energy of $\mathbf{x}$ is contained in a small number of the elements of $\mathbf{X}$.

At the sensor, we take M non-adaptive linear measurements of $\mathbf{x}$, where $K < M < N$, resulting in the $M \times 1$ vector $\mathbf{y}$. This measurement process can be written as

$$\begin{aligned} \mathbf{y} \; &= \; \mathbf{\Phi x} \\ &= \; \mathbf{\Phi \Psi X}, \end{aligned}$$

where $\mathbf{\Phi}_i$ is an $M \times N$ matrix representing the measurement process. For compressed sensing to work, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ must be *incoherent*. Incoherent means that no element of $\mathbf{\Phi}$ has a sparse representation in terms of the elements of $\mathbf{\Psi}$. In order to provide incoherence that is independent of the basis used for reconstruction, a matrix with elements chosen in some random manner is generally used. Thus unlike $\mathbf{\Psi}$, which is constant, $\mathbf{\Phi}$ will change every frame.

Once $\mathbf{y}$ has been obtained, it is transmitted in some fashion to a processor, where it is processed by a reconstruction algorithm. Reconstruction of a compressed sensed signal involves trying to recover the sparse vector $\mathbf{X}$. It has been shown [1, 2] that

$$\hat{\mathbf{X}} \; = \; \arg\min \|\mathbf{X}\|_1 \qquad \text{s.t.} \qquad \mathbf{y} = \mathbf{\Phi\Psi X}, \tag{2}$$

will recover $\mathbf{X}$ with high probability if enough measurements are taken. In general, the ℓ_p norm is defined as

$$\|\mathbf{a}\|_p \; = \; \left(\sum_j |a_j|^p \right)^{\frac{1}{p}}.$$

Equation (2) can be reformulated as

$$\hat{\mathbf{X}} \; = \; \arg\min_{\mathbf{X}} \|\mathbf{y} - \mathbf{\Phi\Psi X}\|_2 \quad \text{s.t.} \quad \|\mathbf{X}\|_0 = K, \tag{3}$$

where the ℓ_0 norm, $\|\mathbf{a}\|_0$ just counts the number of non-zero elements in $\mathbf{a}$.

There are a variety of algorithms to perform the reconstructions in (2) and (3), in this paper we make use of orthogonal matching pursuit (OMP) [14] to solve (3). OMP is a relatively-efficient iterative algorithm that produces one component of $\hat{\mathbf{X}}$ each iteration, and thus allows for simple control of the sparsity of $\hat{\mathbf{X}}$. As the true sparsity is often unknown, the OMP algorithm is run for a pre-determined number of iterations, K, resulting in $\hat{\mathbf{X}}$ being K-sparse.

3. SPARSELY-EXCITED SPEECH

The speech model used in [5] is based on the Nyquist-sampled speech sample sequence $x(n)$ being represented by the convolution relation

$$x[n] \; = \; h[n] * r[n], \tag{4}$$

where $h[n]$ is the signal domain impulse response of the smooth spectral envelope (which in this paper is represented using the Linear Prediction Coefficients - LPC), and $r[n]$ is the residual excitation component. The convolution relation of (4) can be expressed in frame-by-frame matrix form as

$$\mathbf{x} \; = \; \mathbf{hr}, \tag{5}$$

where $\mathbf{h}$ is an $N \times N$ impulse response matrix and $\mathbf{r}$ is an $N \times 1$ excitation vector. We consider linear convolution, and thus $\mathbf{h}$ is Toeplitz lower triangular.

The residual excitation vector is not truly sparse, as for real speech all of the elements of $\mathbf{r}$ will be non-zero. However, the work of [5] showed that $\mathbf{r}$ is indeed highly compressible, and thus (5) is a suitable representation of speech for use with compressed sensing.

So substituting $\mathbf{h}$ and $\mathbf{r}$ into (3) for $\mathbf{\Psi}$ and $\mathbf{X}$ respectively, we obtain

$$\hat{\mathbf{r}} \; = \; \arg\min_{\mathbf{r}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{h}\mathbf{r}\|_2 \quad \text{s.t.} \quad \|\mathbf{r}\|_0 = K \qquad (6)$$

Unfortunately, the basis matrix $\mathbf{h}$ is signal-dependent, and the authors of [5] solve this problem by constructing a codebook of L basis matrices from training speech data. Given that $\mathbf{h}$ is formed by the LPC coefficients of the speech signal, L is in fact the codebook size formed using the LPC vectors, represented as Line Spectral Frequencies (LSFs).

By performing the compressed sensing reconstruction over the codebook, the complexity is linearly scaled by L, and (6) becomes

$$\hat{\mathbf{h}}_l, \hat{\mathbf{r}} \; = \; \arg\min_{\mathbf{h}_l, \mathbf{r}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{h}_l\mathbf{r}\|_2 \quad \text{s.t.} \quad \|\mathbf{r}\|_0 = K, \qquad (7)$$

where $l = 1, 2, \ldots, L$.

4. SPEAKER IDENTIFICATION

4.1 GMM Speaker Identification

As a baseline, we implemented the speaker identification system of [6], which is a simple but powerful system that has been shown to successfully perform this task. This is a GMM-based system, where for each one of the speakers in the database, a corpus is used to train a GMM model of the extracted sequences of (short-time) spectral envelopes. Thus, for a predefined set of speakers a sufficient amount of training data is assumed to be available, and identification is performed based on segmental-level information only. During the identification stage, the spectral vectors of the examined speech waveform are extracted and classified to one of the speakers in the database, according to a maximum *a posteriori* criterion. More specifically, a group of S speakers in the training dataset is represented by S different GMM's $\lambda_1, \lambda_2, \ldots, \lambda_S$, a sequence (or segment) of n consecutive spectral vectors $X = [\mathbf{x}_1\mathbf{x}_2 \cdots \mathbf{x}_n]$ is identified as spoken by speaker $\hat{s}$ based on:

$$\hat{s} = \arg\max_{1 \le q \le S} p(\lambda_q|X) = \arg\max_{1 \le q \le S} \frac{p(X|\lambda_q)p(\lambda_q)}{p(X)}. \qquad (8)$$

For equally likely speakers and since $p(X)$ is the same for all speaker models the above equation becomes

$$\hat{s} = \arg\max_{1 \le q \le S} p(X|\lambda_q), \qquad (9)$$

and finally, for independent observations and using logarithms, the identification criterion becomes

$$\hat{s} = \arg\max_{1 \le q \le S} \sum_{k=1}^{n} \log p(\mathbf{x}_k|\lambda_q), \qquad (10)$$

where

$$p(\mathbf{x}_k|\lambda_q) = \sum_{i=1}^{M} p_q(\omega_i)\mathcal{N}(\mathbf{x}_k; \boldsymbol{\mu}_{i,q}, \mathbf{\Sigma}_{i,q}), \qquad (11)$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$ denotes a Gaussian density with mean $\boldsymbol{\mu}_i$ and covariance $\mathbf{\Sigma}_i$. Note that this is a text-independent system, *i.e.* the sentences during the validation stage need not be the same as the ones used for training. As in [6], the error measure employed is the percentage of segments of the speech recording that were identified as spoken by the most likely speaker. A segment in this case is defined as a time-interval of pre-specified duration containing n spectral vectors, during which these vectors are collectively classified based on (10), to one of the speakers by the identification system. If each segment contains n vectors (n depending on the pre-specified duration of each segment), different segments overlap as shown below, where Segment #1 and Segment #2 are depicted:

$$\overbrace{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n}^{\text{Segment \#1}}, \mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \cdots$$

$$\mathbf{x}_1, \overbrace{\mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1}}^{\text{Segment \#2}}, \mathbf{x}_{n+2}, \cdots$$

The resulting percentages are an intuitive measure of the performance of the system. There is a performance decrease when decreasing the segment duration, which is an expected result since the more data available, the better the performance of the system. A large number of segments is also important for obtaining more accurate results; it should be noted, though, that an identification decision is made for each different segment, independently of the other segments.

4.2 Speaker Identification Using CS

In order to perform speaker identification using compressed sensing, we propose forming a codebook of basis matrices from speech training data for each of the S speakers that we wish to identify. This is essentially formed by performing a codebook of the LSF vectors of each speaker separately. This process is in fact similar to the GMM training for speaker identification, and is based on the assumption that LSF's are suitable feature vectors for the classification task.

A simple way to do classification using compressed sensing is to find a basis for each of the C classes of interest, and then reconstruct a sparse vector from each of the class bases. The measured signal is then said to come from the class that produced the sparsest recovered vector. This can work well, but requires that the class bases be incoherent.

In our case, the class bases would be the $\mathbf{h}_l$'s for each speaker. Unfortunately these bases are far from incoherent. We thus need to find another method to perform speaker identification, and we proceed in the following manner.

We first find a residual excitation vector for each basis matrix from each speaker's codebook using

$$\hat{\mathbf{r}}_{s,l} \; = \; \arg\min_{\mathbf{r}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{h}_{s,l}\mathbf{r}\|_2 \quad \text{s.t.} \quad \|\mathbf{r}\|_0 = K. \qquad (12)$$

Once these have been found, we then calculate

$$d_s \; = \; \min_{l} \|\mathbf{y} - \mathbf{\Phi}\mathbf{h}_{s,l}\hat{\mathbf{r}}_{s,l}\|_2, \qquad (13)$$

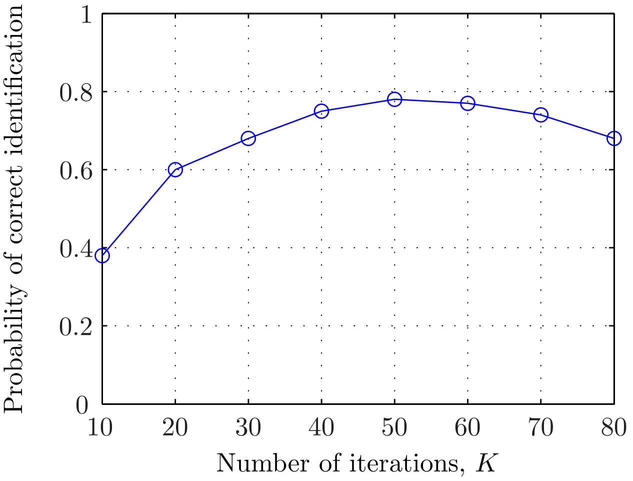

Figure 1: Probability of correct identification versus the number of iterations of the reconstruction algorithm for a codebook size of 16. The number of measurements is equal to half the Nyquist rate ($M = N/2$).

which represents the minimum distance between the measurements $\mathbf{y}$ and measurements from the reconstructions from the s-th speaker's codebook.

Now, let $d_{i,s}$ be the d_s calculated for the i-th frame. The actual speaker s^* in the i-th frame should have the smallest distance, so that

$$d_{i,s^*} \;<\; d_{i,s}, \qquad \forall s \neq s^*. \tag{14}$$

Thus if this is true we have chosen the correct speaker, and if not we have an error.

In practice, we can greatly improve the reliability of speaker identification by considering n frames at a time (*i.e.* a segment as defined in Section 4.1). This is based on the fact that the speaker will not change from frame to frame, and will rather be constant for a group of frames. Thus we use a sliding window to determine the most probable speaker as

$$\hat{s} \;=\; \arg\min_{s} \sum_{j=i-(n-1)}^{i} d_{i,s}, \tag{15}$$

to determine the speaker. Obviously if $\hat{s} \neq s^*$ then the identification has failed for this particular segment. This approach is the same as the segment-based approach for identification as explained in the last part of Section 4.1.

5. RESULTS AND DISCUSSION

We now discuss the results we obtained from the simulation of our proposed method. All speech signals used in training and testing were obtained from the VOICES corpus, available by OGI's CSLU [15]. The speech signals, originally sampled at 22 kHz, were downsampled to 8kHz, with $N = 320$ samples per frame and 50% overlapping between frames. The training data consisted of 30 sentences from 12 speakers, resulting in around 6000 frames per speaker.

Our codebooks were constructed in a manner similar to that of [5], we analysed the training data to obtain the LPC and LSF vectors from which we generated the set of $\mathbf{h}_l$'s for each speaker. We chose to use an LPC order of 22 as that provided better performance, and results in no increase in run-time complexity.

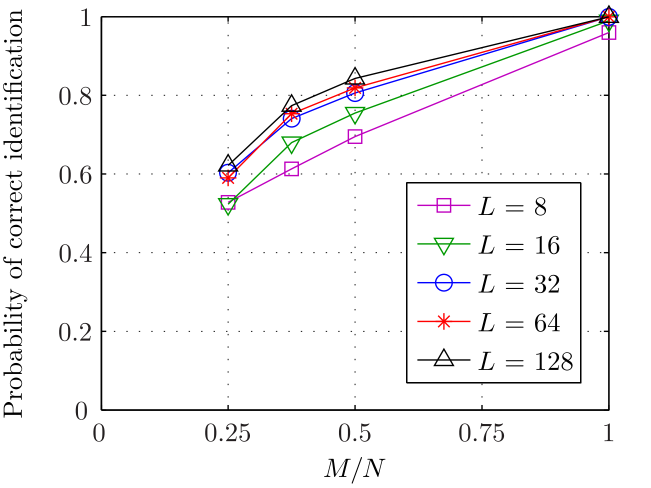

Figure 2: Probability of correct identification versus the number of measurements for various speaker codebook sizes. The number of iterations of the reconstruction algorithm is equal to one quarter of the number of measurements ($K = M/4$).

All the simulations were performed using 10 sentences for each speaker *different* to those used to generate the codebooks. This provided more than 2000 frames of test data for each speaker.

Initially, we tested the performance of the sparsity-based speaker identification. The measurement matrix $\mathbf{\Phi}$ consisted of $M \times N$ Gaussian samples with zero mean and unit variance. The performance measure used was the probability of correct identification of the speaker using (15) with n equal to 140 frames (2.8 seconds), and averaged over all 12 speakers.

As an initial investigation, we looked at the effect of the number of iterations of the OMP algorithm, K, on our proposed method for $M = N/2$ measurements per frame. The results are shown in Fig. 1 for a codebook size of $L = 16$. The identification process can be seen to not be very sensitive to K around $K \approx M/4$, and it is this value for K that we used in the rest of this work.

Fig. 2 presents the performance of our proposed method as the number of measurements M and the size of each speaker's codebook L are varied. These results are intuitively satisfying; as M decreases, the reconstruction quality will degrade, and thus the probability of correct identification decreases. The results for $M/N = 1$ do not use compressed sensing, and this can be thought of as the best possible performance. The performance also improves as L increases, although there seems to be diminishing returns after $L = 32$, and each increase in L increases the complexity of the identification process. Thus for $L = 32$ with 50% measurements the probability of correct identification is about 0.8, and if the measurements are lowered to 25% this probability drops to about 0.6.

All the previous results are for noise-free speech.

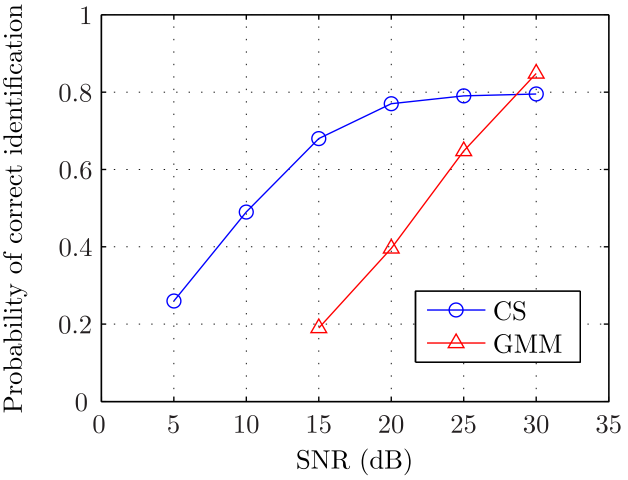

Figure 3: Probability of correct identification versus the signal to noise ratio of the speech signal for the Gaussian mixture model (GMM) method and the proposed compressed sensing (CS) method.

We also explored the effect of additive white Gaussian noise on the probability of correct identification for the $L = 32$, $M = N/2$ case, and this is presented in Fig. 3, along with the corresponding results for the GMM method discussed in Section 4.1. The GMM method used 32 diagonal mixtures and the same training and testing data as the compressed sensing (CS) method. It is clear that the CS method outperforms the GMM method once the signal to noise ratio (SNR) is below 30dB. In fact, there is very little loss in performance for the CS method down to an SNR of 20dB, and even an SNR of 15dB affects the performance mildly.

Assuming the two methods were used in a sensor with limited power resources, the CS method would require slightly more processing than the GMM method in the sensor, as it needs to calculate the measurements, although efficient measurement methods do exist. However the CS method would require half the bandwidth of that of the GMM method to transmit the measurements back to a central processor.

This transmission power gain and the robustness to noise for the CS method come at the cost of increased complexity in the speaker identification algorithm, but for many applications this is acceptable.

6. CONCLUSIONS

We have presented a novel method for speaker identification based on a sparse signal model and the use of compressed sensing. The use of compressed sensing permits the use of less transmission power for the sensor recording the voice. Additionally, our method has been shown to be robust to noise in the recorded speech signal. This is encouraging and warrants further investigation.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[3] R.G. Baraniuk, "Compressive sensing," *IEEE Sig. Proc. Mag.*, pp. 118–120, July 2007.

[4] A. Griffin, T. Hirvonen, A. Mouchtaris, and P. Tsakalides, "Encoding the sinusoidal model of an audio signal using compressed sensing," in *Proc. IEEE Int. Conf. on Multimedia Engineering (ICME'09), New York, NY, USA*, June 2009.

[5] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan*, 2009, pp. 4125–4128.

[6] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3(1), pp. 72–83, January 1995.

[7] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: Applications to speech coding based on sparse linear prediction," *IEEE Signal Proc. Lett.*, vol. 17, no. 1, pp. 103–106, January 2010.

[8] K. H. Yuo, T. H. Hwang, and H. C. Wang, "Combination of autocorrelation-based features and projection measure technique for speaker identification," *IEEE Trans. on Speech and Audio Processing*, vol. 13(4), pp. 565–574, July 2005.

[9] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Speaker identification under noisy environments by using harmonic structure extraction and reliable frame weighting," in *in Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH)*, Pittsburgh, Pennsylvania, USA, September 2006.

[10] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16(6), pp. 1097–1111, August 2008.

[11] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, April 2008.

[12] K. Kumar, Q. Wu, Y. Wang, and M. Savvides, "Noise robust speaker identification using Bhattacaryya distance in adapted gaussian models space," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, August 2008.

[13] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(5), pp. 1711–1723, July 2007.

[14] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.

[15] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.