MULTIPLE SOUND SOURCE LOCALIZATION BASED ON TDOA CLUSTERING AND MULTI-PATH MATCHING PURSUIT

Hong Liu, Bing Yang, Cheng Pang

Key Laboratory of Machine Perception,
Shenzhen Graduate School, Peking University
hongliu@pku.edu.cn, {bingyang, chengpang}@sz.pku.edu.cn

ABSTRACT

Multiple sound source localization in wireless acoustic sensor networks (WASNs) is a challenging problem. Although compressive sensing based methods have shown effectiveness in uncorrelated sources localization, their performance degrades significantly when they are used to locate multiple speech sources. To this end, we propose a multiple sound source localization method based on the time difference of arrival (TDOA) clustering and the multi-path matching pursuit algorithm. First, TDOAs are calculated locally in time-frequency (TF) bins of sensor recordings. Then, the TDOAs are clustered after utilizing outlier rejection to remove erroneous estimations. Finally, a multi-path matching pursuit algorithm is proposed to solve a sparse localization model for localizing multiple sound sources. Experimental results show that the proposed method yields good performance for multiple sound source localization, especially in strong noisy scenarios.

Index Terms— Multiple sound source localization, TDOA clustering, multi-path matching pursuit, wireless acoustic sensor networks

1. INTRODUCTION

Multiple sound source localization is an essential issue for many signal processing tasks, such as echo cancellation, blind source separation and noise reduction [1, 2]. Although many related works have achieved good performance by using fixed microphone arrays [3–5], multiple sound source localization remains a challenge for sources located in a large field due to the limited spatial sampling ability of conventional arrays. With the great flexibility in sensor placement [6], wireless acoustic sensor networks (WASNs) can physically cover a large area, thus showing great potential for multiple sound source localization [7], especially for sound sources distributed in a large space.

This work is supported by National High Level Talent Special Support Program, National Natural Science Foundation of China (NSFC, No.61340046, 61673030, U1613209), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011), Natural Science Foundation of Guangdong Province (No.2015A030311034), Scientific Research Project of Guangdong Province (No.2015B010919004).

Several different methods have been proposed to estimate source location based on time difference of arrival (TDOA) [8–11], direction of arrival (DOA) [4, 5, 12] and received signal strength (RSS) [13–16]. In [2], a Bayesian clustering based approach is proposed to estimate sound source location. However, the sensor node is constrained in a fixed circular microphone array and it lacks flexibility in sensor placement. Some authors focused on time-frequency (TF) processing [4, 5]. They exploited the non-stationary and sparse property of audio signals in TF domain to improve the robustness in noisy and multi-source scenarios, yet only a few works extend these traditional microphone TF methods to WASNs. Recently compressive sensing (CS) [17] has been used to localize multiple sources in wireless sensor networks [11, 13–15]. Though many CS-based approaches have shown great effectiveness, most of them have strict constrains on source signals, such as known energy and uncorrelated source signals, which is not suitable for speech signals. Therefore, their performance degrades significantly when locating speech sources. To get rid of the limitation of existing CS-based methods, a novel approach based on the sparse localization model [11] is proposed to localize both uncorrelated sources and multiple speech sources in WASNs.

In this paper, we propose a multiple sound source localization approach based on TDOA clustering and multi-path matching pursuit (MPMP). First, the generalized cross correlation-phase transform (GCC-PHAT) is utilized to compute TDOAs locally in each TF bin. In this way, the sparse property of speech signal is taken into consideration to improve robustness to noise and multi-source influence. Then, the outlier rejection is adopted to remove erroneous TDOA estimations for depressing their negative effects on clustering. After that, the final TDOA measurements are estimated by K-means clustering, in which way the statistical information of TDOAs is employed. Finally, the MPMP algorithm is proposed to recover source position based on the sparse localization model. The localization method achieves good performance in strong noisy environments, since it takes full advantage of the sparse property of speech signals by TF processing and multiple search paths are added to find the

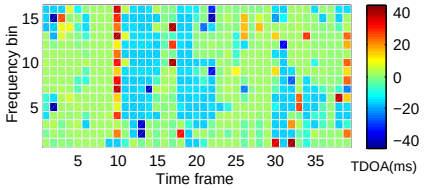


Fig. 1. Estimated TDOAs in TF bins for two sound sources in strong noisy environment (SNR = 0dB). Different colors indicate different TDOA values.

optimized location in the novel MPMP algorithm.

2. TDOA CLUSTERING

Since speech signals are not uncorrelated white sequences, extracting TDOAs directly by finding maximum peaks of cross-correlation function [11] is not suitable. Therefore, the sparse property of speech signals that at most one source is dominant in each TF bin is taken into consideration.

2.1. TDOA Estimation in TF Domain

There are M synchronized microphones capturing source signals in a WASN with only one microphone in each sensor node. One reference sensor is chosen from these M microphones. The reference sensor and each of the other microphones will form a microphone pair. For each microphone pair, detailed processing procedure is shown as follows.

The signals captured by each microphone pair are divided into overlapping time frames. For each frame, the short-time Fourier transform (STFT) coefficients are computed and then they are utilized to derive the cross-spectrum. By using a PHAT-weighted [18] cross-spectrum, the cross-correlation function for each TF bin is calculated as:

$$C_{n,l}(m) = F^{-1}\left\{\frac{1}{|S_{n,l}(\omega)|}S_{n,l}(\omega)\right\}, \tag{1}$$

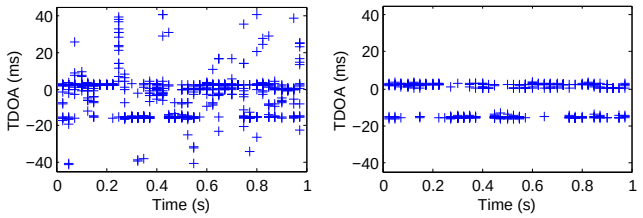
where n represents the time frame index, l denotes the frequency bin index, ω is the frequency point index in l-th frequency bin which contains a series of frequency-adjacent points, $S_{n,l}(\omega)$ denotes cross-spectrum of signals received by each sensor pair, $1/|S_{n,l}(\omega)|$ is the PHAT weight for $S_{n,l}(\omega)$, $C_{n,l}(m)$ is the cross-correlation function with m being time sample point, and $F^{-1}\{\cdot\}$ represents the inverse STFT. With the cross-correlation function $C_{n,l}(m)$, the TDOA $\tau_{n,l}$ for n-th frame and l-th frequency bin can be estimated as:

$$\tau_{n,l} = \underset{m}{\operatorname{argmax}}\left\{C_{n,l}(m)\right\}. \tag{2}$$

In this way, multiple TDOA estimates $\{\tau_{n,l}\}$ in TF domain are derived. Fig. 1 shows an instance in which sensor recordings with the length of one second are divided into 39 frames and 16 frequency bins. It can be seen that two different TDOA groups (colored in green and blue) cover most bins, indicating that there are two dominant sources.

2.2. Outlier Rejection and Clustering

As shown in Fig. 2(a), there are some inaccurate TDOA estimates with large deviations, which have negative effect on



(a) (b)

Fig. 2. The effect of outlier rejection for two active sources in strong noisy environment (SNR = 0dB). (a) The distribution of TDOAs before outlier rejection. (b) The distribution of TDOAs after outlier rejection.

TDOA clustering. Thus the outlier rejection [2] is utilized to remove these low-cardinality erroneous TDOAs from the set of estimates $\{\tau_{n,l}\}$. After that, the remaining estimate of $\{\tau_{n,l}\}$ is denoted by $\{\Gamma_q\}$ with $q \in \{1,\ldots,Q\}$. Fig. 2(b) shows the results processed by the outlier rejection.

Assume there are K active sound sources. The K-means clustering algorithm is applied to find an optimal partition that assigns Q estimates to a set of clusters $\mathbb{S} = \{\mathbb{S}_1,\ldots,\mathbb{S}_K\}$. The K-means algorithm assigns each estimate to the closest cluster, and each cluster centroid can be updated as the average of all estimates in this cluster. After several iterations, the optimal assignment can be obtained by minimizing the following objective function [19]:

$$J(\mathbb{S}) = \sum_{k=1}^{K}\sum_{\Gamma_q\in\mathbb{S}_k}||\Gamma_q - \Delta^{(k)}||^2, \tag{3}$$

where $\mathbb{S}_k$ denotes the set of estimates that belongs to k-th cluster with $k \in \{1,\ldots,K\}$, $\Delta^{(k)}$ is the centroid of $\mathbb{S}_k$, and $||\cdot||$ represents the Euclidean norm.

With the optimal assignment, K measurements can be obtained by averaging all assigned estimates in each TDOA cluster. After all the sensor signals are transmitted to the fusion center of the WASN, TDOA measurements for all the microphone pairs are computed using the above-mentioned method. TDOA measurements corresponding to j-th microphone and the reference sensor are represented by $\{\hat{\Delta}_j^{(t)}\}$ with $j \in \{1,\ldots,M-1\}$.

3. MULTI-PATH MATCHING PURSUIT BASED SPARSE LOCALIZATION

3.1. Sparse Localization Model

With TDOA measurements $\{\hat{\Delta}_j^{(t)}\}$, a sparse localization model is utilized to estimate the source location. By using summed version TDOA measurements, the model avoids the data association problem [12] which exists in direct matching approaches. The summed version TDOA measurements are constructed by summing TDOAs and the absolute values of TDOAs respectively:

$$y_j = \sum_{t=1}^{K}\hat{\Delta}_j^{(t)},\ y'_j = \sum_{t=1}^{K}|\hat{\Delta}_j^{(t)}|, \tag{4}$$

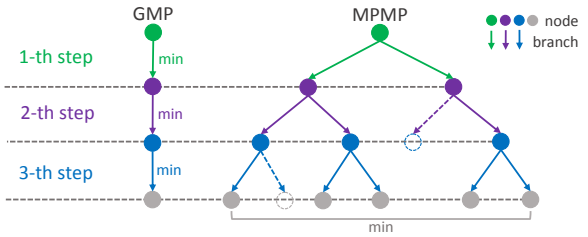


Fig. 3. Example of the search trees for GMP and MPMP algorithm. The number of sources is set to three. The dash lines colored in blue and purple denote the erroneous branches. With two branches for each node, five valid search paths are generated by MPMP.

where y_j denotes the summed version of TDOA measurements for j-th sensor, and y'_j represents the summed version of measurements modulus for j-th sensor.

A WASN is distributed in a two-dimension area, which is averagely divided into N discrete grids. M microphones and K sound sources ($K \ll N$) are randomly distributed in different grids and each grid contains at most one sound source. Considering the source sparsity in this area, the source localization problem can be cast into a sparse representation framework [11] expressed as:

$$\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{s} + \boldsymbol{\epsilon}, \tag{5}$$

with

$$\boldsymbol{y} = [y_1, \ldots, y_{M-1}, y'_1, \ldots, y'_{M-1}]^T,$$
$$\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_{2(M-1)}]^T,$$
$$\boldsymbol{s} = [s_1, \ldots, s_N]^T,$$
$$\boldsymbol{\Psi} = \begin{bmatrix} \Delta_{11} & \cdots & \Delta_{1N} \\ \vdots & \ddots & \vdots \\ \Delta_{(M-1)1} & \cdots & \Delta_{(M-1)N} \\ |\Delta_{11}| & \cdots & |\Delta_{1N}| \\ \vdots & \ddots & \vdots \\ |\Delta_{(M-1)1}| & \cdots & |\Delta_{(M-1)N}| \end{bmatrix},$$

where $\boldsymbol{y}$ denotes the measurement vector, $\boldsymbol{\epsilon}$ contains the additive noise $\boldsymbol{\epsilon}_j$ on TDOA measurements for j-th sensor. Here, $\boldsymbol{s}$ is the source vector, and s_i denotes the number of sound sources in i-th grid with $i \in \{1, \ldots, N\}$ and $s_i \in \{0, 1\}$. $\boldsymbol{\Psi}$ represents TDOA fingerprinting matrix, in which Δ_{ji} is the theoretical TDOA value for j-th microphone from a sound source located in i-th grid. The source vector contains all-zero elements except for K elements that correspond to the grids where the sound sources locate. Therefore, the location can be estimated by solving the K-sparse problem.

3.2. Multi-Path Matching Pursuit Algorithm

There are many compressive sensing (CS) reconstruction algorithms [20] available to recover the sparse vector $\boldsymbol{s}$. It has been proved that the greedy matching pursuit (GMP) algorithm is more superior in source localization compared with other popular CS algorithms [15]. Based on the GMP, we propose a novel multi-path matching pursuit (MPMP) algorithm, which searches multiple paths to find the optimal location of each sound source.

Algorithm 1: Proposed MPMP algorithm

Input: Fingerprint matrix $\boldsymbol{\Psi}$, measurement vector $\boldsymbol{y}$, the number of sound sources K, the number of branches B

Output: Estimated source vector $\hat{\boldsymbol{s}}$

```
T ← 0        → Initialization of search tree
P ← 1        → Initialization of the number of search paths in T
for k = 1, ..., K do
    for p = 1, ..., P do
        compute all the cost for each grid using Eq. (6)
        select B branches corresponding to B minimum cost and remove the erroneous branches
        add branches to search paths in T
    end for
    P ← the number of search paths in T
    update residual measurement vector for P paths
end for
seek the optimal search path in T that minimizes the Euclidean norm of residual measurement vector
compute ŝ according to the optimal search path of T
return ŝ
```

In order to find accurate grids for K sound sources, K steps are performed in the MPMP algorithm. As shown in Fig. 3, the grid is selected according to the cost function at each step. The cost function I of selecting one grid from the area is defined as:

$$I = ||\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{z}||, \tag{6}$$

where the temporal source vector $\boldsymbol{z} = [z_1, \ldots, z_i, \ldots, z_N]^T$. Here, $\boldsymbol{z}$ contains all-zero elements except $z_i = 1$ with $i \in \{1, \ldots, N\}$ denoting that only one sound source locates at i-th grid. $\boldsymbol{r}$ is an $N \times 1$ residual measurement vector which is initialized as $\boldsymbol{r} = \boldsymbol{y}$. $\boldsymbol{\Psi}\boldsymbol{z}$ indicates the contribution of i-th grid to $\boldsymbol{r}$.

As depicted in Fig. 3, the GMP algorithm has only one search path. At each step, the grid with the minimum cost is selected and the contribution of the selected grid is subtracted from $\boldsymbol{r}$. However, with the influence of noise, the minimum-cost grid is not exactly the true source location, which leads to a biased residual measurement for the following search steps. Compared with the GMP, multiple search paths are adopted in our MPMP algorithm as depicted in Fig. 3. For each step, B branches corresponding to B minimum cost are selected for each node, which forms the search tree T for localizing multiple sound sources. Fig. 3 presents an instance of $B = 2$. In practice, the erroneous branches whose cost is larger than $||\boldsymbol{r}||$ are removed. For each branch, the contribution of selected grid is subtracted from the residual measurement to update $\boldsymbol{r}$. Finally, the optimal sound source location is indicated by the path with the minimum $||\boldsymbol{r}||$. The GMP can be seen as a case

Table I. The RMSEs of MPMP and GMP algorithm for different SNRs.

SNR	-10dB			0dB			10dB			20dB		
The number of sources	1	2	3	1	2	3	1	2	3	1	2	3
MPMP	**0.00**	**0.66**	**1.49**	**0.00**	**0.59**	**1.47**	**0.00**	**0.58**	**1.42**	**0.00**	**0.56**	**1.36**
GMP	0.00	0.93	1.66	0.00	0.84	1.65	0.00	0.83	1.60	0.00	0.81	1.57

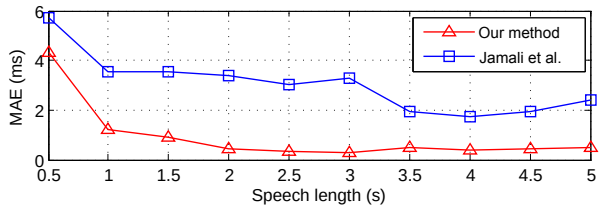


Fig. 4. Performance evaluation for TDOA estimation with respect to different speech lengths (SNR = 0dB).

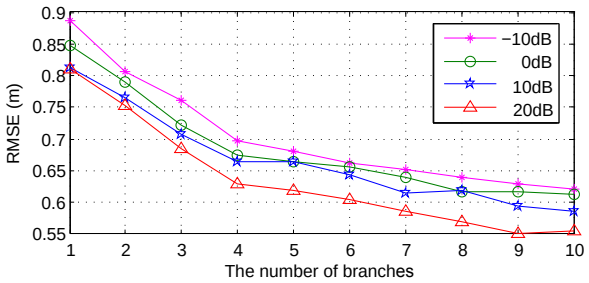


Fig. 5. Performance evaluation for MPMP algorithm with different numbers of branches.

of the MPMP with $B = 1$. The GMP identifies the source gird in an order that the source in the minimum-cost grid is selected first, and it can only provide a local-optimum solution. The MPMP is not limited by this order, because more paths are added and the optimum path is chosen from multiple candidate search paths. The details of proposed MPMP is described in Algorithm 1.

4. EXPERIMENTS AND ANALYSIS

We simulated a 10m $\times$ 10m area averagely partitioned into 20 $\times$ 20 = 400 square grids. Assume 20 microphone nodes and up to 3 simultaneous sound sources are randomly deployed [13] at different grid centers. Speech recordings sampled at 16kHz are selected from the TIMIT database [21]. The SNR is measured as the ratio of the received source signal power to noise power at sensor location. In each experiment, the frame length is set to 800 with 50% overlap. The STFT size is set to 1600, and the STFT coefficients are equally divided into 16 frequency bins with 100 frequency points in each bin. The experimental results are calculated as the mean of 100 independent Monte Carlo (MC) runs.

Assume that there are two active sound sources in the area with SNR = 0dB. We utilize the mean absolute error (MAE) to evaluate the performance of TDOA estimation. The MAE is computed by averaging all the absolute errors between the elements of $\boldsymbol{y}$ and corresponding theoretical values. Fig. 4 shows the MAEs with different speech lengths for our method and the TDOA estimation approach proposed by Jamali *et al*. [11]. It can be seen that the proposed method achieves smaller MAEs than [11], because with multiple TDOA estimates in TF bins the statistical information is made full use of in our approach. Besides, the estimation error of both methods degrades sharply when the speech length is increased from 0.5s to 1s. When the speech length is long enough, the MAEs fluctuate slightly.

In the following experiments, speech recordings with the mean length of one second are utilized. The accuracy of localization is evaluated by the root mean square error (RMSE) between the estimated positions and true positions. To assess MPMP algorithm, the RMSEs for different branch numbers B is shown in Fig. 5. The presented RMSE values are the average of the cases when the number of sound sources is one, two and three. When the number of branches increases, the RMSEs reduce at the cost of the increased processing time. Fig. 5 shows that the RMSEs degrade significantly when the number of branches is less than 4. Considering the complexity and accuracy of this localization algorithm, the number of branches B is set to 4 in the following experiments.

The GMP algorithm has been illustrated to be superior to other popular CS algorithms in source localization [11]. Therefore, only the GMP algorithm is compared with proposed MPMP algorithm in this experiment. Table I depicts the RMSEs for the MPMP and GMP algorithm in the environments where there are one, two and three sound sources with different SNRs respectively. When there is one sound source, both algorithms have similar performance because they work in the same way to find one grid that contributes most to the measurement vector $\boldsymbol{y}$. Since that TDOAs are estimated accurately without multi-source influence, both methods perform well. When two or three sound sources exist in the scenario, it is evident that the MPMP outperforms the GMP with lower RMSEs for different SNRs, due to that MPMP algorithm searches multiple paths for optimal location.

5. CONCLUSIONS

In this paper, we propose an effective approach for multiple sound source localization in WASNs based on TDOA clustering and multi-path matching pursuit. By taking full use of the sparse property of speech signals, source signals is free of strict constrains as traditional compressive sensing based source localization methods described. The TDOA clustering based on outlier rejection and K-means achieves accurate and noise-robustness TDOA estimation for multiple sound sources. The multi-path matching pursuit algorithm is applied to sparse localization model, which can achieve the optimal location estimation by adding search paths.

6. REFERENCES

[1] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1079–1093, 2016.

[2] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation and counting in a wireless acoustic sensor network," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015.

[3] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4771–4783, 2015.

[4] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[5] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2287–2291, 2014.

[6] J. Zhang, R. C. Hendriks, and R. Heusdens, "Structured total least squares based internal delay estimation for distributed microphone auto-localization," *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2016.

[7] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," *IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, pp. 1–6, 2011.

[8] D. Ayllon, R. Gil-Pita, M. Rosa-Zurera, and H. Krim, "Real-time multiple doa estimation of speech sources in wireless acoustic sensor networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2709–2713, 2015.

[9] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 439–443, 2013.

[10] L. Yang and K. C. Ho, "An approximately efficient TDOA localization algorithm in closed-form for locating multiple disjoint sources with erroneous sensor positions," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4598–4615, 2009.

[11] H. Jamali-Rad and G. Leus, "Sparsity-aware TDOA localization of multiple sources," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4021–4025, 2013.

[12] A. Griffin, A. Alexandridis, D. Pavlidi, and A. Mouchtaris, "Real-time localization of multiple audio sources in a wireless acoustic sensor network," *European Signal Processing Conference (EUSIPCO)*, pp. 306–310, 2014.

[13] E. Lagunas, S.K. Sharma, and S. Chatzinotas, "Compressive sensing based target counting and localization exploiting joint sparsity," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3231–3235, 2016.

[14] H. Jamali-Rad, H. Ramezani, and G. Leus, "Sparsity-aware multi-source RSS localization," *Signal Processing*, vol. 101, pp. 174–191, 2014.

[15] B. Zhang, X. Cheng, N. Zhang, Y. Cui, Y. Li, and Q. Liang, "Sparse target counting and localization in sensor networks based on compressive sensing," *IEEE International Conference on Computer Communications (INFOCOM)*, pp. 2255–2263, 2011.

[16] X. Sheng and Y. H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 44–53, 2005.

[17] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[19] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[20] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium*, 1993.