Hierarchical classification of speaker and background noise and estimation of SNR using sparse representation

K V Vijay Girish [1], *A G Ramakrishnan*[1] *and T V Ananthapadmanabha*[2]

[1]Indian Institute of Science, Bangalore, India
[2] Voice and Speech Systems, Bangalore, India
kv@ee.iisc.ernet.in, ramkiag@ee.iisc.ernet.in, tva.blr@gmail.com

Abstract

In the analysis of recordings of conversations, one of the motivations is to be able to identify the nature of background noise as a means of identifying the possible geographical location of a speaker. In a high noise environment, to minimize manual analysis of the recording, it is also desirable to automatically locate only the segments of the recording, which contain speech. The next task is to identify if the speech is from one of the known people. A dictionary learning and block sparsity based source recovery approach has been used to estimate the SNR of a noisy speech recording, simulated at different SNRs using ten different noise sources. Given a test utterance, a noise label is assigned using block sparsity approach, and subsequently, the speaker is classified using sum of weights recovered from the concatenation of speaker dictionaries and the identified noise source dictionary. Using the dictionaries of the identified speaker and noise sources, framewise speech and noise energy are estimated using a source recovery method. The energy estimates are then used to identify the segments, where speech is present. We obtain 100% accuracy for background classification and around 90% for speaker classification at a SNR of 10 dB.

Index Terms: noise source, speaker, classification, dictionary, ASNA, segmental SNR, detection, speech segments

1. Introduction

Real life speech signals generally contain foreground speech by a particular speaker in the presence of a background environment like factory or traffic noise. In this work, we address classification of the speaker and background noise source, and then frame-wise energy estimation of the audio sources. Identification of background noise can help us to narrow down to possible geographical locations of the speaker. Apriori estimate of speaker and the background noise is useful for speech enhancement, separation and speech recognition; which has been of common interest to research community and finds many applications in the real world. Frame-wise energy estimation of speech source is useful for identifying speech segments. If this is possible, then the low SNR recordings can be automatically processed to extract only the speech regions. These speech segments can then be processed by human experts, in defense applications. Other applications are hearing aids [1], forensics [2] and robotic navigation systems [3]. In this work, we address the problem of speaker as well as noise classification of noisy speech signals using the concept of block sparsity [4] and sparse non-negative recovery [5]. We also estimate the segmental SNR and detect the speech segments in a noisy speech signal.

1.1. Literature review

There has been a lot of work on audio content analysis and scene classification. Lu et. al. [6] classified audio into speech, music, environment sound and silence using K-nearest neighbor and line spectral pairs-vector quantization. Barchiesi et.al. [7] presents a review of the state of the art in acoustic scene classification. Giannoulis et al. [8] evaluated 11 algorithms along with a baseline system for acoustic scene classification. The algorithms extracted time and frequency domain features from the audio signal followed by a statistical model or majority vote based classifier. Cauchi [9] classified auditory scenes using non-negative matrix factorization.

Machinery noise diagnostics was explored in [10] while acoustic signature classification of aircrafts or vehicles was surveyed in [11]. Noise was classified for hearing aid applications based on variation of signal envelope as features in Kates [12]. Line spectral frequencies were used as features for classification of different kinds of noise and speech by Maleh et. al. [13]. Casey [14] devised a system using a hidden Markov model classifier and log-spectral features to classify twenty different types of sounds. Chu et al. [15] combined matching pursuit based features with mel-frequency cepstral coefficients to recognize 14 different environmental sounds. Chachada et. al. [16] surveyed techniques for stationary and non-stationary environmental sound recognition.

Tzagkarakis et. al. [17] used sparsity based speaker identification using discriminative dictionary learning while Joder et. al. [18] explored non-negative matrix factorization for feature extraction. Malkin [19] explored machine listening research to solve real world problems in perceptual computing. Our paper addresses some components of machine listening like scene and speaker identification, and SNR estimation.

Kim et. al. [20] estimated the SNR based on the analysis of waveform amplitude distribution. Narayanan et. al. [21] estimated SNR based on computational auditory scene analysis. Liu et. al. [22] estimated the SNR of clipped signals based on the audio amplitude distribution. Tchorz et. al. [23] estimated SNR in different frequency channels using amplitude modulation spectrograms.

Dictionary learning is a method of representing features from a large training data using a weighted linear combination of vectors called as atoms. Estimating weights corresponding to these atoms is termed as sparse coding or source recovery. Audio signals are represented as a linear combination of non-negative dictionary atoms for audio source separation [24, 25, 26], recognition [27, 28, 29], classification [30, 31] and coding [32, 33]. The simplest dictionary learning (DL) method is a random selection of features from the training data

[5]. Matching pursuit [34], orthogonal matching pursuit (OMP) [35], focal underdetermined system solver (FOCUSS) [36] and basis pursuit [37] are some of the source recovery algorithms.

In our work, we have used active-set Newton algorithm (ASNA)[5] for source recovery. The training phase for the classification problem is DL from various speaker/ noise sources. The dictionary atoms capture the variation in the spectral characteristics of the speech and noise sources.

1.2. Contributions

The major contributions of this paper are: (1) Block sparsity and concatenated dictionary based classification of speech and noise sources from noisy speech (2) Exploiting low and high energy segments of noisy speech signal for noise and speaker class estimation (3) A novel algorithm to detect speech segments and measures to quantify the segment detection and segmental SNR.

2. Proposed Approach

2.1. Problem definition

Noisy speech signal, $s[n]$ is simulated as a linear combination of two sources, a speech, $s_{sp}[n]$ and a noise source, $s_{ns}[n]$.

$$s[n] = s_{sp}[n] + s_{ns}[n] \tag{1}$$

The speech and noise are constrained to belong to a specific set of speakers and noise sources, and the test signal is classified as belonging to one of the predefined speaker and noise sources. Figure 1 shows a part of an utterance from a female speaker, factory noise and the noisy speech signal at an SNR of 0 dB .

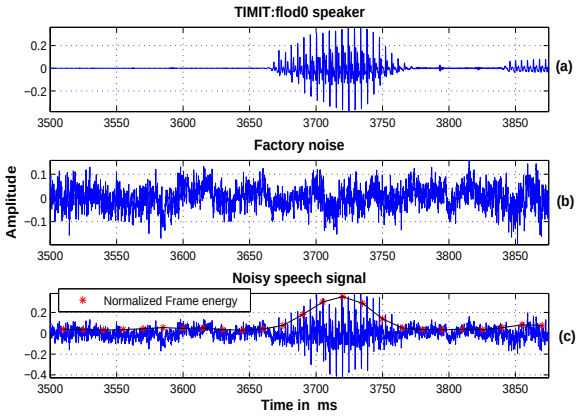

Figure 1: Illustration of speech, noise and the noisy speech signal. Star marks in (c) indicate frame-wise energies.

2.2. Feature extraction and dictionary learning

From the training set of speech and noise sources, frames of 60 ms duration are extracted with a shift of 15 ms. The magnitude of the short-time Fourier transform of these frames are used as the features. A dictionary is a matrix $D \in \mathbb{R}^{\mathrm{P}\times\mathrm{K}}$ containing K column vectors denoted as atoms, $d_k,\ 1 \le k \le K$. A given feature vector y can be represented as a linear combination of a few dictionary atoms as $y \approx Dx$, where $x \in \mathbb{R}^{\mathrm{K}}$ is the vector of weights for the atoms. Dictionary is learnt by random selection of $K(= 500)$ features as atoms. Features for each speaker and noise source are extracted separately and the corresponding dictionaries are built. The dictionaries for N_{sp} speaker and N_{ns} noise sources are denoted as $D^i_{sp},\ 1 \le i \le N_{sp}$ and $D^i_{ns},\ 1 \le i \le N_{ns}$, respectively.

2.3. Classification using block sparsity and source recovery of the mixed signal

Since the test signal is assumed to have at most two sources, the features extracted can be approximated as a linear combination of atoms belonging to two source dictionaries. Let y be the test feature containing the n^{th} speaker and the m^{th} noise source, which can be represented as

$$y \approx \hat{y} = [D^n_{sp}\ D^m_{ns}][x'_n\ x'_m]' \tag{2}$$

where n, m and the weight vectors x_n, x_m are unknown, and are estimated. Estimation of n, m simultaneously is difficult as there are $N_{sp} \times N_{ns}$ combinations of audio sources, which can form the mixed signal. We estimate the noise source first, and then the speech source as explained below.

2.3.1. Noise classification stage

The background noise source is normally stationary in nature, while the speech foreground is non-stationary. Also the speech component consists of voiced, unvoiced and silence segments. So, when speech and noise sources are mixed at a particular SNR, the frames containing silence segments of speech are the ones with least energy since they contain only the noise components as illustrated in Fig.1(c). These low energy frames give a higher confidence in the estimation of the noise class. Given l frames from $s[n]$, and the corresponding features $y_i,\ 1 \le i \le l$, the energy of each frame is $E_y(i) = ||y_i||_2^2$. Ten features having the lowest $E_y(i)$ are extracted as $Y_{min} = [y_{(1)} \ldots y_{(10)}]$. A concatenated dictionary is constructed from the individual noise source dictionaries as $D_{ns} = [D^1_{ns}..D^{N_{ns}}_{ns}]$, and let Y_{min} correspond to the m^{th} noise source. The j^{th} column in Y_{min} can be represented as

$$y_{(j)} \approx [D^1_{ns}..D^{N_{ns}}_{ns}][x'_1 \ldots x'_{N_{ns}}]' \tag{3}$$

where $x = [x'_1 \ldots x'_{N_{ns}}]'$ is block-sparse [4, 38], i.e. $||x_m||_2$ is non-zero and $||x_i||_2 = 0,\ \forall i \neq m$, D^m_{ns} is the dictionary block corresponding to x_m having non-zero Euclidean norm. The noise source is estimated as the index $\hat{m}$, which gives the maximum absolute sum of correlation between the dictionary atoms and the ten features (similar to block-OMP [38]):

$$\hat{m} = \arg\max_i \sum_{j=1}^{10} ||(D^i_{ns})^T y_{(j)}||_1 \tag{4}$$

2.3.2. Speaker classification stage

To estimate the speaker index, we approximate the test feature y_i as the linear combination of the dictionary atoms from the estimated noise source, $D^{\hat{m}}_{ns}$ and the concatenated dictionary of speech sources $[D^1_{sp} \ldots D^{N_{sp}}_{sp}]$. Since speech comprises silence, unvoiced and voiced regions, we use only those features having higher energy (60% of the total number of features) for speaker classification. Since speech is non-stationary and noise further corrupts it, the speaker source index is determined by comparing the weights estimated in the representation:

$$y \approx [D^1_{sp} \ldots D^{N_{sp}}_{sp} D^{\hat{m}}_{ns}][x'_1 \ldots x'_{N_{sp}} x'_{\hat{m}}]' = Dx \tag{5}$$

and the weight vector, x is estimated by minimizing the distance $dist(y, Dx)$ using ASNA [5], where $dist()$ is the KL-divergence between y and Dx.

$$\underset{\mathbf{x}}{\text{minimize}}\ KL(y||\hat{y}),\ \hat{y} = Dx\ s.t.\ x \geq 0 \tag{6}$$

The weight vector x recovered using ASNA is non-negative and sparse. A new measure *Total Sum of Weights (TSW)* is defined as the total absolute sum of elements of $x_i,\ 1 \leq i \leq N_{sp}$ for all the selected features y_j,

$$TSW_i = \sum_j ||x_i||_1, \forall y = y_j,\ 1 \leq j \leq l \tag{7}$$

$$\hat{n} = \arg\max TSW_i \tag{8}$$

TSW_i is observed as the reliable measure for estimating the speaker source index, $\hat{n}$ as the x recovered using ASNA [5] tends to pick up atoms corresponding to the original source dictionary D^n_{sp} and $D^{\hat{m}}_{ns}$ iteratively, and also assigns higher weights for D^n_{sp}. Figure 2 shows the plot of sum of weights over all the features for each speaker source dictionary. It is seen that the highest sum of weights is obtained for the original speaker fl0d0, except for the SNR of -10 dB.

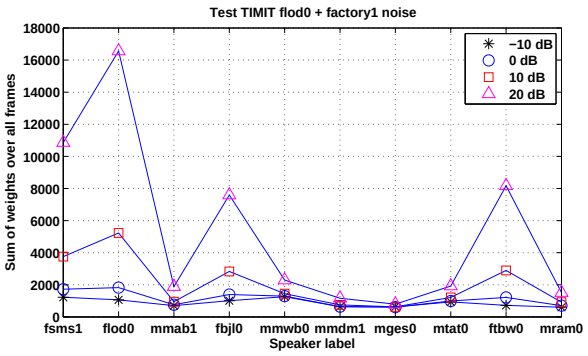

Figure 2: Weights estimated for each speaker source.

2.4. Estimation of SNR and the speech segments

Using the estimated speaker and noise source indices, $\hat{m}, \hat{n}$, the speech and noise components of the features corresponding to the noisy speech signal are recovered by using a concatenated dictionary, $D = [D^{\hat{n}}_{sp}\ D^{\hat{m}}_{ns}]$ and recovery algorithm ASNA similar to eqn.(6). So the estimated feature $\hat{y}$ and estimates of speech, $\hat{y}_{sp}$ and noise features, $\hat{y}_{ns}$ are

$$\hat{y} = [D^{\hat{n}}_{sp}\ D^{\hat{m}}_{ns}][x'_{\hat{n}}\ y'_{\hat{m}}]' \tag{9}$$

$$\hat{y}_{sp} = D^{\hat{n}}_{sp} x_{\hat{n}},\ \hat{y}_{ns} = D^{\hat{m}}_{ns} x_{\hat{m}} \tag{10}$$

Segmental SNR is defined as the ratio of the total energy of the speech to noise features in decibel, for segments in the signal, where speech is present. Estimates of frame-wise SNR, SNR^i_{fr} for the i^{th} frame and the segmental SNR, SNR_{seg} are defined as

$$SNR^i_{fr} = 10 \log \frac{||y^i_{sp}||^2_2}{||y^i_{ns}||^2_2} \tag{11}$$

$$SNR_{seg} = 10 \log \frac{\sum_i ||y^i_{sp}||^2_2}{\sum_i ||y^i_{ns}||^2_2} \tag{12}$$

Figure 3 shows the plot of frame-wise energies for the original, estimated and noisy speech features, and also indicates the frames, where speech segments are present, for an SNR of 0 dB. It is seen from the figure that high energy frames actually contain speech. The reasoning for this is that speech definitely has voiced regions, which have high energy. Also the voiced frames are interspersed with unvoiced and silence frames. So, the local maxima in the speech region corresponding to voiced frames are significantly higher than the local maxima in the frames corresponding to noise segments. This demarcation between the maxima in speech and noise frames helps to extract the speech segments by using k-means clustering algorithm to extract the significant maxima and the corresponding speech region around the same.

We propose a novel algorithm for extracting the speech segments as shown in Algorithm 1. The number of clusters, k can be increased so as to get more distinct segments, which is to be handled in the future. Figure 3 shows the local maxima and the two clusters, where it can be seen that black squares (elements of cluster 2) are a subset of the speech frames.

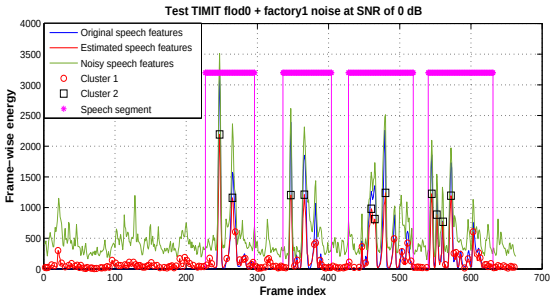

Figure 3: Frame-wise energy of original and estimated speech.

Algorithm 1 *Detection of the speech segments*

1: Pick the local maxima of the frame energies.
2: Do k-means clustering of the local maxima with k=2 and initialize centroids as the maximum and minimum among the local maxima.
3: Pick the cluster elements with higher centroid value and assign them as $cluster_{max}$ which correspond to the voiced segments of the speech.

end

We define measures Miss Rate and False alarm rate with respect to the detection of the speech segment :

- Miss rate (MR): Percentage of number of speech segments, which do not encompass any element of $cluster_{max}$ with respect to the total number of speech segments.
- False alarm rate (FAR): Percentage of number of $cluster_{max}$ which are outside the speech segments with respect to the total number of $cluster_{max}$.

3. Results and discussion

The database for speech sources is taken from randomly selected ten speakers from dialect 5 of the training set of the TIMIT database. For each speaker, 8 utterances are used for

training and the rest 2, for testing. The duration of each utterance is 2-4 seconds. The ten noise sources are taken from NOISEX database [39]. The first 10 seconds of each noise source is used for testing and the rest is used for training. Test utterances for each speaker are equally divided into two segments so that we get four speech segments. These segments are added to the noise test signal at randomly selected locations ensuring a minimum of 200 ms gap between successive speech segments. So, the test signal consists of noise signal interspersed with speech segments added at segmental SNR of $-10,\ 0,\ 10$ and 20 dB.

Table 1 shows the estimated sources for all the combinations of speaker and noise sources at a SNR of 0 dB. It is seen that the speaker sources in the presence of white noise are misclassified the most, followed by the factory and jet cockpit noise.

Table 1: Confusion matrix showing the estimate of speaker and noise sources for all the combinations of speaker and noise sources at a SNR of 0 dB. $*$ marked cells are combinations which are correctly classified. All noise sources are correctly classified, so only misclassified speakers are shown in the table. All speakers are correctly classified in the presence of babble, car interior, tank, military vehicle and destroyer operations noises, which are not shown.

Noise > Speaker	*white*	*factory1*	*hfchnl*	*f16*	*jet*
fsms1	*	*	*	*	*
flod0	fbjl0	*	*	*	*
mmab1	fbjl0	mmwb0	*	mtat0	mtat0
fbjl0	*	fsms1	fsms1	*	fsms1
mmwb0	*	*	*	*	*
mmdm1	*	*	*	*	*
mges0	fbjl0	mmwb0	*	mtat0	mmwb0
mtat0	fbjl0	*	*	*	*
ftbw0	fbjl0	*	*	*	*
mram0	fbjl0	mmwb0	mges0	*	*

Note- f16: f16 cockpit, jet: jet cockpit, hfchnl: high frequency channel

Table 2 shows the overall classification accuracies for speaker and noise sources for SNR of -10 to 20 dB. 100% accuracy is achieved for noise sources at all SNR since we exploit the low energy frames of the mixed signal to estimate the noise class. Speaker classification accuracy is above 80% for SNR above 0 dB but degrades at lower SNRs. Using higher energy frames instead of all the frames of the test signal for speaker classification increased the accuracy significantly from 62% to 83% at 0 dB SNR. Joder et. al. [18] reported a speaker classification accuracy of 98.9 % for eight speakers with clean speech, while we achieve 100% accuracy at 20 dB SNR. Figure 4 shows the plot of mean absolute value and standard deviation of the error between the estimated segmental SNR and actual SNR averaged over all the combinations of noise with speech segments from each speaker. The overall mean absolute and standard deviation of the error is 1.39 and 1.08 at 0 dB SNR. The error in the estimated SNR is due to the noise component being represented by the atoms of speech dictionary or vice versa. Loizou [40] performed speech enhancement and evaluated improvement in segmental SNR in speech with speech-shaped noise while we estimate the original segmental SNR in the noisy speech signal.

Table 2: Classification accuracy of speaker and noise sources

SNR (dB)	-10	0	10	20
Speaker (%)	37	83	99	100
Noise (%)	100	100	100	100

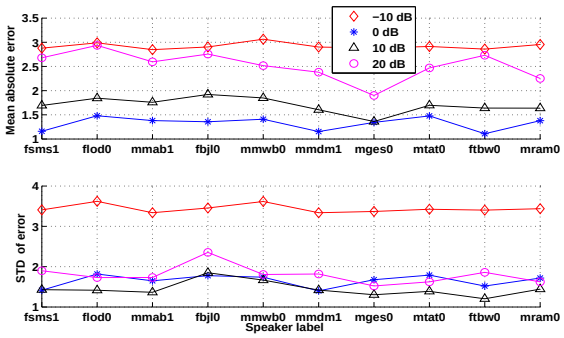

Figure 4: Mean absolute and standard deviation of error between the original and the estimated segmental SNR for each speaker, for different simulated SNRs.

Table 3 shows the variation of miss and false alarm rates for the detection of speech segments. We get a miss rate of 0.75% and FAR of 0.17% at a SNR of 0 dB. Wak et. al. [41] studied the voice activity detection (VAD) techniques and performed spectral subtraction for speech enhancement before using energy based VAD. Fukuda et. al. [42] proposed a statistical model based noise robust VAD algorithm, where they reported speech segment detection rate of 95% averaged over high and low SNR while we report average miss rate of 3.25% for SNR of -10, 0 and 10 dB (the measures used are not exactly the same).

Table 3: Miss and false alarm rates in speech segment detection

SNR (dB)	-10	0	10	20
MR (%)	8.50	0.75	0.50	0.50
FAR (%)	8.84	0.17	0.00	0.00

4. Conclusion and future work

We have shown speaker and noise classification from noisy speech signals with good classification accuracy using a simple dictionary learning method and sparse representation. We have also shown SNR estimation and detection of speech segments with very low miss and false alarm rate at various SNRs. We plan to explore other dictionary and discriminative learning methods. Also, the numbers of speakers and noise sources can be increased so as to test the scalability of the proposed approach. Further, the adaptation of dictionaries in the case of an unknown source is planned as a future work.

5. Acknowledgement

We are grateful to Defence Research and Development Organization, Govt. of India for funding the project.

6. References

[1] R Turner, [Online] *http://www.wired.co.uk/news/archive/2013-10/02/machine-hearing-cambridge-university*

[2] S. Ikram, " Digital audio forensics using background noise," *Multimedia and Expo (ICME)*, July 2010, pp. 106-110.

[3] S. Chu, S. Narayanan, C. C. Jay Kuo, and M. J. Matari, "Where am I? Scene recognition for mobile robots using audio features," *In IEEE International Conference on Multimedia and Expo*, pp. 885-888, 2006.

[4] Y. C. Eldar, P. Kuppinger and H. Bolcskei, "Block-Sparse Signals: Uncertainty Relations and Efficient Recovery," *IEEE Trans. Signal. Process.*, vol. 58, no-6, pp. 3042 - 3054, 2010.

[5] T. Virtanen, J. F. Gemmeke, B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, pp. 2277 - 2289, 2013.

[6] L. Lu and H.Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 7, 2002.

[7] D. Barchiesi, D. Giannoulis, D. Stowell, M. D. Plumbley and P. Mermelstein, "Acoustic Scene Classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[8] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange and M. D. Plumbley, "Detection and classification of acoustic scenes and events: an IEEE AASP challenge," *IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, Oct. 2013.

[9] B. Cauchi, "Non-negative matrix factorization applied to auditory scene classification," *Masters thesis, ATIAM (UPMC / IRCAM / TELECOM ParisTech)*, 2011.

[10] R. H. Lyon, "Machinery Noise and Diagnostics," *Butterworth-Heinemann*, 1987.

[11] A. Shirkhodaie, and A. Alkilani, "A survey on acoustic signature recognition and classification techniques for persistent surveillance systems," *Proc. Signal Processing, Sensor Fusion, and Target Recognition*, May 2012.

[12] J. M. Kates, "Classification of background noises for hearing aid applications," *J. Acoust. Soc. Am.*, vol. 91 (1), pp. 461-470, Jan. 1995.

[13] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," *Proc. IEEE Conf. Acoustics, Speech, Signal Proc*, March 1999, pp. 237-240.

[14] M. Casey, "Reduced-rank spectra and minimum-entropy priors as consistent and reliable cues for generalized sound recognition," *Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis, Eurospeech*, Aalborg, Denmark, 2001.

[15] S. Chu, S. Narayanan and C. C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol.17, no.6, 2009.

[16] S. Chachada and C. C. J. Kuo, " Environmental sound recognition: A survey," *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2013, p. 1–9.

[17] C. Tzagkarakis and A. Mouchtaris, "Sparsity based robust speaker identification using a discriminative dictionary learning approach," *Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.

[18] C. Joder and B. Schuller "Exploring Nonnegative Matrix Factorization for Audio Classication: Application to Speaker Recognition," *Proceedings of Speech Communication*, 2012, pp. 1–4.

[19] R. G. Malkin, "Machine listening for context-aware computing," *Doctoral Dissertation, Carnegie Mellon University*, 2006

[20] C. Kim and R. M. Stern, "Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis," *Interspeech*, 2008,

[21] A. Narayanan and D. Wang, "A CASA-Based System for Long-Term SNR Estimation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, pp. 2518-2527, 2012.

[22] X. Liu, J. Jia and L. Cai, "SNR Estimation for Clipped Audio Based on Amplitude Distribution." *International Conference on Natural Computation (ICNC)*, 2013, pp. 1434-1438.

[23] J. Tchorz and B. Kollmeier, "SNR Estimation Based on Amplitude Modulation Analysis With Applications to Noise Suppression," *IEEE Trans. Speech and Audio Process.*, vol. 11, 2003.

[24] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol.15, no.3, 2007.

[25] A. Ozerov, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 3, 2010.

[26] G. J. Mysore, P. Smaragdis, and B. Raj, " Non-negative Hidden Markov Modeling of Audio with Application to Source Separation," *Lecture Notes in Computer Science, Latent Variable Analysis and Signal Separation*, vol. 7572, pp. 186-199, 2012.

[27] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 3, 2010.

[28] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 7, 2011.

[29] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," *Interspeech 2010*, Tokyo, Japan, 2010.

[30] Y. C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classication," *Pattern Recognition Letters*, vol. 26 (9), 2005.

[31] S. Zubair, F. Yan, W. Wang "Dictionary learning based sparse coefficients for audio classification with max and average pooling," *Elsevier Digital Signal Processing* , vol. 23, issue. 3, 2013.

[32] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," *Proceedings of the 128th Audio Engineering Society Convention*, London, UK, 2010.

[33] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proceedings of the IEEE,* vol. 98 (6), pp. 995-1005, 2009.

[34] S. G. Mallat, and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Sig. Process.*, vol. 41, pp. 3397-3415, 1993.

[35] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 1993.

[36] I. F. Gorodnitsky, and B. D. Rao, Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm, *IEEE Trans. Sig. Process.*, vol. 45, pp. 600-616, 1997.

[37] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43 (1), pp. 129-159, 2001.

[38] Y. Fu, H. Li, Q. Zhang and J. Zou, "Block-sparse recovery via redundant block OMP," *Signal Processing, Elsevier*, vol. 97, pp. 162–171, 2014.

[39] Noisex-92. [Online], Available: *http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html*

[40] P. C. Loizou, "Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 857–869, 2005.

[41] M. W. Mak and H. B. Yu, " A study of voice activity detection techniques for NIST speaker recognition evaluations," *Elsevier Computer Speech and Language*, vol. 28, pp. 295–313, 2014.

[42] T. Fukuda, O. Ichikawa and M. Nishimura, " Long-Term Spectro-Temporal and Static Harmonic Features for Voice Activity Detection," *IEEE Journal of Selected Topics in Sig. Process.*, vol. 4, pp. 834–844, 2010.