Noise-robust feature based on sparse representation for speaker recognition

Hongzhuo Qi

School of Computer Science, Harbin University of Science and Technology,
Harbin 150080, Heilongjiang, China

Abstract
The performance of speaker recognition suffers substantial degradation in noisy environments. To solve this problem, we propose a new feature extraction method based on sparse coding to improve the noise robustness of the speaker recognition. In this method, an over-complete dictionary, which is powerful in identifying transient underlying structures and harmonic periodicities, is trained with amounts of unlabeled data. Next, speech signals are sparsely represented by atoms of the dictionary. After that, the proposed feature is extracted from the sparse representation by a tuning function which simulates the hearing mechanism. Experiments show that the proposed method outperforms the mel-frequency cepstral coefficients (MFCC) and the perceptual linear predictive (PLP) feature in various noise conditions.

Key words: UNDERLYING STRUCTURE, AUDITORY TUNING, SPARSE REPRESENTATION, ROBUST FEATURE, SPEAKER RECOGNITION

Introduction

Although current speaker recognition systems can achieve satisfied accuracy rates, their performances are degraded substantially by environment noises. The main reason is that current speech feature not only represents the speech, but also represents the noises, resulting in a mismatch between training and testing. Currently, two widely used features are mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) [1]. MFCCs mainly simulate the properties of the auditory system, while LPCCs model the vocal tract with an all-pole model. These two methods can achieve satisfied performance under ideal environments. Another widely used feature is the perceptual linear predictive (PLP) feature, which also simulate the mechanism of hearing. However, their performances drop rapidly in adverse condition. Even some de-noise techniques are used; the performance is still degraded because current speech enhancement methods inevitable cause further distortion. Therefore, it is an essential issue to improve the robustness of speaker recognition system by extracting noise-robust features.

Time-relative and non-stationary transient underlying structures such as time relations between acoustic events and harmonic periodicities provide important cues for recognition. The Fourier transform is not suitable to capture these cues, since it is not only cumbersome to represent such transient phenomenon, but also sensitivity to noise [2]. That is why the features based on Fourier transform are sensitive to noise. It is found that the sparse representation can capture the underlying time-frequency structures of sounds at low-bit rates [3]. Therefore, we need a new technique to improve the robustness of speech features.

Current researches show that the neurosensory systems encode stimuli by activating only a small number of neurons out of a large population at the same time [4, 5]. This indicates that the sensory perceptual system of human beings may process signals in a sparse manner. Further studies show that many natural signals, such as image and speech, are sparse or approximately sparse [6]. Sparseness seems to be an inherent property of signals and can be treated as a prior knowledge. Sparse coding makes use of such knowledge and therefore provides effective representations of signals. Through years of intensive research, sparse coding has been applied successfully in image processing [7]. Recently, speech processing algorithms, which achieve improved performance by using sparse models [8] -[11], encourage researchers to make more and further explorations.

Further proofs shows that over-complete representations have strong robustness to noise [12]. On the basis of those researches, two types of robust features are proposed in the literature. The first type is extracted by sparse representation plus Gammatone filters, and its robustness to noise is confirmed by experimental results [13]. The second type is a binary feature generated by the sparse representation over a learned over-complete dictionary. Experiments show that this feature is noise-robust and discriminative [14]. In this paper, we introduce an auditory tuning function to replace 0-norm function to extract a novel noise-robust feature. Experiments show that the 0-norm function is a special case of the modified auditory tuning function.

The proposed methods can be summarized as follows. Firstly, an over-complete dictionary is learned from speech corpus, such that the atoms can be more efficient to represent the underlying structures of the speech. Then, speech signals are sparsely represented over the learned dictionary and the sparse representations are obtained. Next, the sparse representations are transformed into robust features by a modified auditory tuning function [15]. Finally, a new speaker model is designed to capture the speaker-specific information and classify speech signals. Experiments show that the proposed feature is robust to noise and discriminative between different speakers.

Over-complete representation

Over-complete dictionary learning

The question of the first importance in sparse coding is dictionary preparation. The dictionary can either be chosen as a predefined set of functions (such as steerable wavelets, curvelets, contourlets) or be designed by adapting its content to fit a given set of signal observations [16]-[18]. Current researches show that a dictionary by learning with amounts of data can perform better than a dictionary by choosing. In order to represent the underlying structures, an optimal over-complete dictionary is learned from a speech database. The dictionary learning can be cast as the following optimization problem

$$\arg\min_{\Psi,C}\|C\|_0 + \lambda\|X-\Psi C\|_2^2 \qquad (1)$$

where $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1(t) \dots \boldsymbol{\psi}_L(t)] \in R^{K\times L}$ denotes an over-complete dictionary which is initialized by cosine and sine functions; $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N] \in R^{K\times N}$ is a speech frame set with each column $\boldsymbol{x}_n$ being a K-dimension speech frame; $\boldsymbol{C} = [\boldsymbol{c}_1 \dots \boldsymbol{c}_N] \in R^{L\times N}$ is the sparse coefficient set of $\boldsymbol{X}$; $\|.\|_F$ denotes the l_F-norm; and λ is a regular parameter which is proportional to residual energy level. Note that each function $\boldsymbol{\psi}_l(t)$ of $\boldsymbol{\Psi}$ is a unit norm function:

$$\int_{-\infty}^{+\infty}\boldsymbol{\psi}_l^2(t)dt = 1 \qquad (2)$$

Since there are two problems: both of the l_0-norm and the joint optimization of Ψ and C are non-convex functions, sothe formulaa (1) is difficult to conduct (NP-Hardproblemem). Fortunately, Candes et.al [19] have proved that l_0-norm in equation (2) can be replaced by l_1-norm if the signal is sparse. In addition, the joint optimization problem can be simplified by alternating optimization between Ψ and C. So the optimization problem in (1) can be rewritten as:

$$\begin{cases}\arg\min_{C}\|C\|_1 + \lambda\|X-\Psi C\|_2^2\\ \arg\min_{\Psi}\|C\|_1 + \lambda\|X-\Psi C\|_2^2\end{cases} \qquad (3)$$

In the learned dictionary, the number of base functions is more than the dimensionality of the input signal vectors. If the dictionary is well learned, it can be maximized information transfer with the lowest bit rate when signals have strong correlations with the underlying structures [12].

Sparse representation

The auditory pathway is sensitive to acoustic underlying structures, and tends to maximize the encoded information with the minimum acoustic underlying structures. Its code mode is similar to the property of sparse representation [20].

In order to find the "best matching impulses", Lasso algorithm [21] is used to decompose acoustic signals into sparse coefficients in terms of the aforementioned learned over-complete dictionary. For a given speech frame x_n $(n = 1, \dots, N)$ and the learned over-complete dictionary Ψ, the sparse representation c_n can be obtained by solving the following problem:

$$\arg\min_{c_n}\|c_n\|_1 + \lambda\|x_n - \Psi c_n\|_2 \qquad (4)$$

where $c_n \in R^L$ is the sparse coefficient vector of x_n over Ψ, and c_n satisfies

$$\|c_n\|_0 \ll L. \qquad (5)$$

Each non-zero entries of c_n implies that a certain underlying structure is contained in signal x_n, and the underlying structure can be perceived by auditory pathway. In auditory pathway, each non-zero entry represents a neuron impulse, and the amplitude of this impulse will be tuned by a tuning function before transmitting to the auditory cortex.

Speech feature and speaker model

Feature extraction method

Researchers suggested that the tuning function plays a key role in the human auditory system, and the hearing will be lost [15] without the tuning function. In order to simulate the auditory processing mode, the tuning function is introduced to transform the sparse coefficients.

With the learned over-complete dictionary and Lasso algorithm, each input speech vector $x_n, (n = 1, \dots, N)$ is decomposed into a sparse coefficient vector $c_n \in R^L$. The entries of the sparse coefficient vector c_n can be written as $c_{ln}, l = 1, \dots, L$. It is suggested that [22] the following mathematic function is suitable for tuning the sparse coefficient,

$$y_{ln} = |c_{ln}|^{2/3} \qquad (6)$$

where y_{ln} is the tuned value of the lth entry of vector c_n, and $y_n = [y_{1n}, \dots, y_{Ln}] \in R^L$ is a vector which contains the tuned value y_{ln}.

The latest research showed that the real tuning function of the auditory system is sharper than formula (6) [15], so the formula should be changed to model this mechanism. To this end, the above formula can be generalized as following equation:

$$y_{ln} = |c_{ln}|^{2/\rho}, \rho > 0 \qquad (7)$$

Where ρ is a variable used to transform the sparse coefficients of $c_{ln}, l = 1, \dots, L$, and y_{ln} is the transformed value correspond to c_{ln}. With a suitable value of variable ρ, the generated vectors $y_n, n = 1, \dots, N$ can be used as speech features for speaker recognition. This kind of feature is called auditory tuning based coefficients (ATBC).

For the purpose of generating a kind of robustness and discriminative feature, different values of variable ρ are tested, and it is found that the condition $\rho \geq 50$ is preferable for ATBC. We further find that when the tuning variable $\rho \to +\infty$, then formula (7) can be deduced as follows

$$y_{ln} = |c_{ln}|^{2/\rho} \cong \begin{cases} 1, \text{ s.t. } c_{ln} \neq 0 \\ 0, \text{ s.t. } c_{ln} = 0 \end{cases} \qquad (8)$$

Different with this tuning function, You et.al [6] have presented a kind of robust feature which is transformed from sparse coefficient by 0-norm function, and the representation of 0-norm function is

$$\tilde{y}_{ln} = \|c_{ln}\|_0 = \begin{cases} 1, \text{ s.t. } c_{ln} \neq 0 \\ 0, \text{ s.t. } c_{ln} = 0 \end{cases} \qquad (9)$$

The feature actual is a kind of binary feature. Obviously, the function (7) will be equal to the 0-norm function (9) when its variable $\rho \to +\infty$, namely $y_{ln} = \tilde{y}_{ln}$. This implies that the 0-norm function is a special case of the auditory tuning function in equation (7).

Speaker model

The ATBC feature is high-dimensional and sparse. The Gaussian mixture model which is widely used in speaker recognition is not suitable for the ATBC feature. Based on the characteristics of ATBC, we find that it is reasonable to suppose that the expectation of one speaker is different from those of the others. So we use the expectation of ATBC to model speakers, and the representation of the expectation can be written as follows

$$E(Y^s) = \int_{t=-\infty}^{+\infty} y_t^s dt \qquad (10)$$

where $E(.)$ and $y_t^s \in R^L$ denote the expectation operator and ATBC feature of the sth speaker respectively. Here, we name the speaker model as a sparse expectation model (shorten as SEM).

Since each column of the above-mentioned learned over-complete dictionary is statistically independent, meaning that the coefficients are non-correlated, so the equation (10) can be rewritten as

$$E(Y^s) = \frac{1}{N^s} \sum_{n=1}^{N^s} y_n^s \qquad (11)$$

Where N^s denotes the total feature number of the sth speaker, and $y_n^s \in Y^s = [y_1^s, \dots, y_{N^s}^s]$ is ATBC feature.

In order to reduce the effects of speech content, we introduce an expectation version of ‘universal background model’ to strengthen this speaker model.

Scoring method

Scoring method is used to measure the similarity between two speaker models. The cosine distance is simple and effective in measuring the similarity of two expectation vectors, so it is chosen as the scoring method between two speaker models. Given a certain input ATBC feature set y, its expectation model $E(y)$, and the target speaker models $E_s, s = 1, \dots, S$, the representation of the scoring method is:

$$v_s = \langle E_s, E(y) \rangle = {E_s}^T E(y) / \sqrt{\|E_s\|_2^2 \|E(y)\|_2^2} \qquad (12)$$

where T denotes the transpose operation, and the v_s is the distance between input speech and the sth target speaker model. The classification is made via

$$i = \arg\min_{1 \leq s \leq S} v_s. \qquad (13)$$

Experimental Evaluation

In order to evaluate the robustness and discriminative ability of the proposed feature, the

NIST SRE-2003 one-speaker limited-data task data and the Chinese-863 database are used for experiments. In NIST SRE-2003, the evaluation set consisted of 356 target speakers (149 males, 207 females) and 28160 test trials (2215 target trials, 25945 impostor trials). In Chinese-863 corpus, the evaluation set was composed by 110 target speakers (55 males, 55 females) and 26400 test trials (2400 target trials, 24000 impostor trials). In the two evaluation sets, each target speaker has an approximately 2-minute long recording and each test trial is an approximately 10-second-long recording. In additional, the TIMIT database is chosen to learn an over-complete dictionary. To simulate the noisy conditions, three noises, namely the White, Car and Street noises from the Noisex-92 database are added artificially to the test speech of the Chinese-863 database to simulate noise conditions.

The speech signals are split as speech frames with a window length of 20 ms and window shift of 10ms. The ALIZE [23] tools are used to establish a speaker recognition system. The ATBCs are computed with the above method, and the length of a feature vector is 171-dimensional.

System configuration

The MFCC and the PLP are used for comparison. The used feature was 13 MFCCs (c0~c12), appending their first- and second-order derivatives (39-dimensional features). The setup of the PLPs is the same to that of the MFCC. The framework of speaker recognition for these two features is based on the GMM-UBM. The feature proposed in this paper employs the SEM system which is called ATBC-SEM for short.

For the systems of the MFCC and PLP, two gender-dependent UBM-GMMs with 2048 Gaussians are trained for the evaluation of MFCC features. One is built for the NIST SRE-2003 one-speaker limited-data task evaluation, and the UBM is trained on 330 speakers (191 females, 139 males and about 2-minute-long recording for each speaker) of the NIST SRE-2002 database. The other one is built for the robustness evaluation, and the UBM is trained on 90 speakers (45 females, 45 males and approximately 5-minute-long recording for each speaker) from the Chinese-863 database. Cepstral mean subtraction and feature warping [1] are adapted to enhance the feature. A simple three-Gaussians energy-based speech detector is employed to remove the silence (non-speech) frames [24]. The T-Norm is also utilized to improve the recognition performance.

For the ATBC-SEM system, two over-complete dictionaries are learned, each one contains 171 columns. One is trained on TIMIT for the decomposition of NIST SRE-2003 one speaker limited-data recordings; the other one is trained in part of the Chinese-863 database (90 speakers, approximately 5-minute-long recording for each speaker) for the decomposition of the above-mentioned Chinese-863 evaluation set recordings. Then, the ATBC feature is generated by the aforementioned methods. After that, the sparse expectation model is computed. Finally, the cosine distance is used to measure the distance between the input speech and the enrolled speakers.

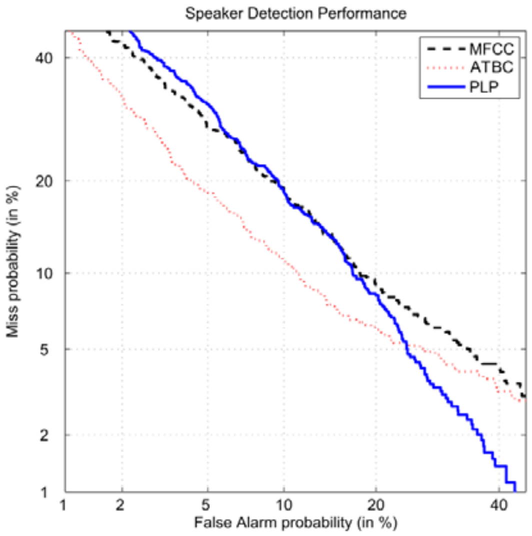

Figure 1. Comparison of the ATBC and MFCC, PLP features on NIST SRE-2003 database

Results and analysis

The DET curve is a convincing indicator in speaker recognition; therefore is chosen to measure the performances. The ATBC-SEM is designed to evaluate the ATBC, and the baseline is used to evaluate the MFCC and the PLP. The results of the evaluation of the NIST SRE-2003 database are shown on Fig. 1. The EER of the MFCC and the PLP are 13.53% and 14.04%, respectively, while that of the ATBC is 10.40%, which is obviously lower the former two features. This result indicates that the ATBC performs better than the MFCC and the PLP under clean conditions.

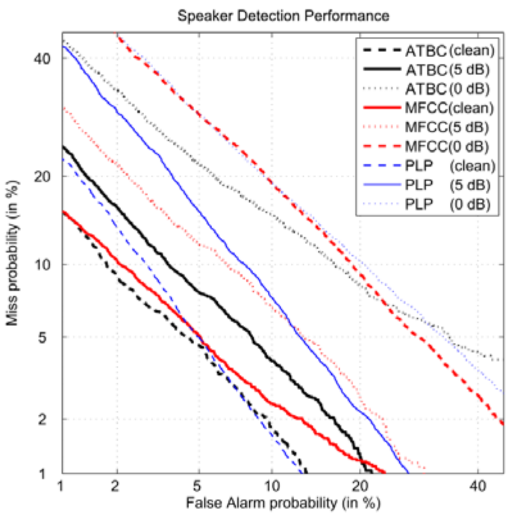

Figure 2. Comparison of the CSC and MFCC, PLP features on Chinese-863 database under clean conditions, and Gaussian White noise condition with SNR=5 and 0 dB.

The results under the Chinese-863 database are shown in Fig. 2. One can see that the EERs of the MFCC and the PLP are 4.94% and 4.86%, respectively. In contrast, the EER of the ATBC is 4.75%. At the SNR of 5 dB, the EERs of the MFCC and the PLP are 8.17% and 8.64%, respectively; the EER of the ATBC is 6.45%. At the SNR of 0 dB, the EERs of the MFCC and the PLP drop down to 14.04% and 14.23%; the EER of the ATBC also drops to 12.69%. These results under different conditions show that the improvement of the ATBC is not obvious under clean condition. With the decrease of the SNR, the ATBC performs much better than the MFCC and the PLP, showing strong noise robustness.

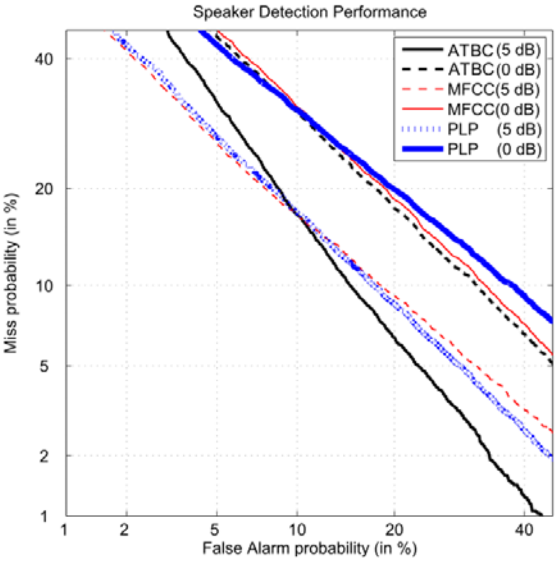

Figure 3. Comparison of the CSC and MFCC, PLP features on Chinese-863 database under Car noise conditions with SNR=5 and 0 dB conditions

The results under the Car noise are shown in Fig. 3. One can see that the EERs of the MFCC are 13.42% at 5dB and 19.35% at 0dB. The PLP performs slightly worse than the MFCC, with EERs of 13.22% at 5dB and 19.69% at 0dB. The ATBC achieves a better performance, yielding EERs of 12.60% and 18.56% respectively. The performances under the Street noise are shown in Fig. 4. A similar result is obtained. The Car and Street noises are time-varying, which raises a challenge to speaker recognition. The above results show that the ATBC always performs better than the MFCC and the PLP.

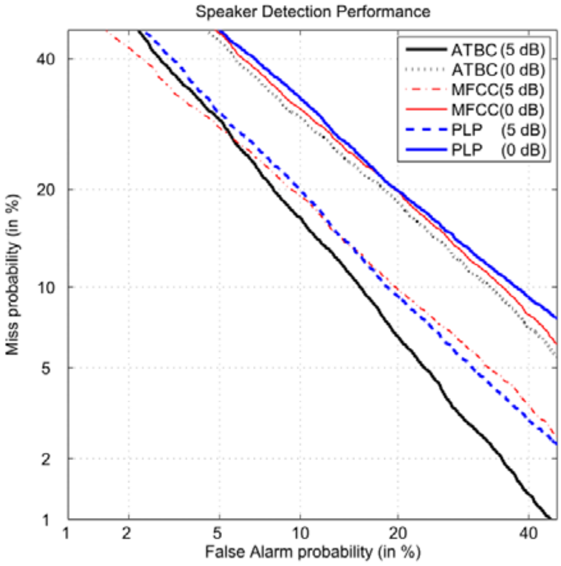

Figure 4. Comparison of the CSC and MFCC, PLP features on Chinese-863 database under Street noise conditions with SNR=5 and 0 dB conditions

Conclusions

In this paper, the underlying structure and auditory tuning based robust feature is presented for speaker recognition. The dictionary which is used to model the underlying structure is learned from speech corpus; and then the input speech is sparsely represented by the learned dictionary; finally, the auditory tuning function is used to transform the underlying structure coefficients into the proposed feature. Experiments show that the feature not only more discriminative but also more noise-robust than MFCC.

Acknowledgements

This work was supported by Youth Scientific Research Foundation Project of Harbin University of Science and Technology (2011YF016).

References

1. Tomi, K., Haizhou, L.(2010) An Overview of Text-independent Speaker Recognition: from Features to Supervector. *Speech Communication*, 52(1), p.p.12-40.

2. Mallat, S. (2008) *A Wavelet Tour of Signal Processing, the Sparse Way*. Academic Press: New York.
3. Smith, E. Lewicki, M.S. (2005) Efficient Coding of Time-relative Structure Using Spikes. *Neural Computation*, 17(1), p.p.19-45.
4. B. Olshausen. (1996) Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, 381(6583), p.p.607-609.
5. D. Attwell, S. Laughlin,(2001) An Energy Budget for Signaling in The Grey Matter of The Brain. *J. Cereb. Blood Flow Metab*., 21(10), p.p.1133-1145.
6. M. A. Davenport, M. F. Duarte, Y. C. Eldar ,G. Kutyniok. (2011) *Compressed sensing: theory and applications*, Cambridge University Press: London.
7. M. Elad, M. A. T. Figueiredo, Yi Ma.(2010)On The Role of Sparse and Redundant Representations in Image Processing. *Proceedings of IEEE*, 98(6), p.p.972-982.
8. T. Virtane. (2007) Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(3), p.p. 1066-1074.
9. C. D. Sigg, T. Dikk, J. M. Buhmann. (2010) Speech Enhancement with Sparse Coding in Learned Dictionaries. *proc. ICASSP*, Dallas, Texas, USA, p.p. 4758-4761.
10. L. L. Durrieu, L. P. Thiran. (2011) Sparse non-negative decomposition of speech power spectra for formant tracking. *proc. ICASSP*, Prague, Czech Republic, p.p. 5260-5263.
11. J. F. Gemmeke, T. Virtanen, A. Hurmalainen. (2011) Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition. *IEEE Trans. on Audio, Speech, and Language Processing*, 19(7), p.p. 2067-2080.
12. Graham, D.J., Field, D.J. (2006) Sparse Coding in the Neocortex. *Evolution of Neuroscience*, 20(5), p.p.887-892.
13. Li, Q., Yan, H. (2010) Robust Speaker Identification Using an Auditory-based Feature. *Proc. ICASSP*, Dallas, Texas, p.p.4515-4517.
14. You, D.T., Jiang, T., Han, J.Q., Zheng, T.R. (2011) A Cochlear Neuron Based Robust Feature for Speaker Recognition. *Proc. ICASSP*, Prague, Czech Republic, p.p. 5440-5443.
15. Reichenbach, Hudspeth, T., A. J. (2010) A Ratch Mechanism for Amplification in Low-frequency Mammalian Hearing. *PANS*, 107(11), p.p. 4973-4978.
16. R. Vidal, Y. Ma, S. Sastry. (2005)Generalized Principal Component Analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12), p.p. 1945-1959.
17. K. Engan, S. O. Aase, J. H. Husoy. (1999) Method of optimal directions for frame design. *Proc. ICASSP*, 5, p.p. 2443-2446.
18. M. Aharon, M. Elad, A. M. Bruckstein. (2006) The K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. on Signal Process*, 54(11), p.p. 4311-4322.
19. Candes, E., Tao, T. (2007) The Dantzig Selector: Statistical Estimation When p is Much Larger Than n. *The Annals of Statistics*, 35(6), p.p. 2313-2351.
20. Smith, E.C., Lewicki, M.S. (2006) Efficient Auditory Coding. *Nature*, 439(7079), p.p. 978-982.
21. Fu, W.J. (1998) Penalized Regressions: the Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics*, 7(3), p.p. 397-416.
22. Barbour, D.L., Wang, X. (2003) Constrast Tuning in Auditory Cortex. *Science*, 299, p.p. 1073-1075.
23. Bonastre, J.F., Wils, F., Meignier, S.(2005) ALIZE, a Free Toolkit for Speaker Recognition. *Proc. ICASSP*, 1, p.p. 737-740.