A PAIRWISE ALGORITHM FOR PITCH ESTIMATION AND SPEECH SEPARATION USING DEEP STACKING NETWORK

Hui Zhang[1], *Xueliang Zhang*[1], *Shuai Nie*[2], *Guanglai Gao*[1], *Wenju Liu*[2]

[1]Computer Science Department, Inner Mongolia University, Hohhot, China, 010021
[2]National Laboratory of Patten Recognition (NLPR), Institute of Automation, University of Chinese Academy of Sciences, Beijing, China, 100190

alzhu.san@163.com, cszxl@imu.edu.cn, nss90221@gmail.com, csggl@imu.edu.cn, lwj@nlpr.ia.ac.cn

ABSTRACT

Pitch information is an important cue for speech separation. However, pitch estimation in noisy condition is also a task as challenging as speech separation. In this paper, we propose a supervised learning architecture which combines these two problems concisely. The proposed algorithm is based on deep stacking network (DSN) which provides a method of stacking simple processing modules in building deep architecture. In the training stage, an ideal binary mask is used as target. The input vector includes the outputs of lower module and frame-level features which consist of spectral and pitch-based features. In the testing stage, each module provides an estimated binary mask which is employed to re-estimate pitch. Then we update the pitch-based features to the next module. This procedure is embedded iteratively in DSN, and we obtain the final separation results from the last module of DSN. Systematic evaluations show that the proposed approach produces high quality estimated binary mask and outperforms recent systems in generalization.

Index Terms— Speech separation, Pitch estimation, Computational auditory scene analysis, Supervised learning

1. INTRODUCTION

In realistic environments, noise usually degrades the speech intelligibility of hearing-impaired listeners or performance of automatic speech recognition (ASR) systems. Speech separation aims to remove noise by separating target speech from background interference. It is helpful for both hearing aids wearers and ASR systems [1, 2]. Computational auditory scene analysis (CASA) is a promising method to solve the speech separation problem [3].

CASA defines the goal of speech separation as computing an ideal binary mask (IBM) [4], which is useful for improving speech intelligibility [5] and the performance of speech/speaker recognition [6, 7]. The IBM is a time-frequency (T-F) mask, which can be computed from premixed target and interference. Specifically, in a T-F unit, if the signal-to-noise ratio (SNR) is greater than a local SNR criterion (LC), the corresponding mask element in the IBM is set to 1 (target-dominant). Otherwise, the mask element is set to 0 (interference-dominant).

When adopting IBM as the computational goal of CASA, we can naturally formulate the speech separation as a binary classification problem [5]. From the viewpoint of classification, the feature selection is important. Many features have been inspected. Those features include: pitch-based features [8], amplitude modulation spectrum (AMS) [9], relative spectral transform and perceptual linear prediction (RASTA-PLP), Mel-frequency cepstral coefficient (MFCC) and Gammatone frequency cepstral coefficient (GFCC) [10] *etc*. Wang *et al.* [10] suggest that *pitch-based features* have a good generalization in speech separation.

Pitch-based features are derived from pitch. But extracting pitch from noisy speech is also a difficult task, especially in low SNR conditions. Generally speaking, on one hand, if the target voice is separated from the background, we can obtain the pitch easily. On the other hand, speech separation performance will get better if pitch estimation is accuracy. Since these two tasks could benefit from each other, speech separation and pitch extraction in noisy conditions are considered to be a *"chicken-and-egg"* problem.

In this paper, we propose a supervised learning system to deal with this *"chicken-and-egg"* problem more concisely.

- Pitch extraction and speech separation are boosted alternately. (Section 2.1)
- Frame-level features are adopted, which consist of spectral features, and pitch-based features. (Section 2.2)
- We use deep stacking network (DSN) to implement our idea of working on the two problems (pitch extraction and speech separation) alternately. (Section 2.3)
- Systematic evaluations show the proposed approach produces high quality estimated binary masks and outperforms recent systems in unmatched noisy conditions. (Section 3)

This research was supported in part by the China National Nature Science Foundation (No.61365006, No.61263037, No.61305027, No.91120303, No.61273267, No.61403370, and No.90820011).

2. SYSTEM DESCRIPTION

2.1. System Overview

The proposed system is illustrated in Fig. 1, which includes training stage and testing stage. In the training stage, premixed speech and noise are utilized to construct the IBM, which is the training target. Here we use frame-level features as input. The features are extracted from mixed signal, which consist of spectral features and pitch-based features (Section 2.2). The classifier is DSN (Section 2.3). To compute pitch-based features, we use ground truth pitch in the training stage. In the testing stage, the pitch we use is estimated iteratively by the proposed method. Specifically, after getting an estimated IBM via the trained classifier, we use it to extract pitch and then update the pitch-based features. The newly updated features are used for the next round speech separation. This iterative process is embedded in DSN (Section 2.4).

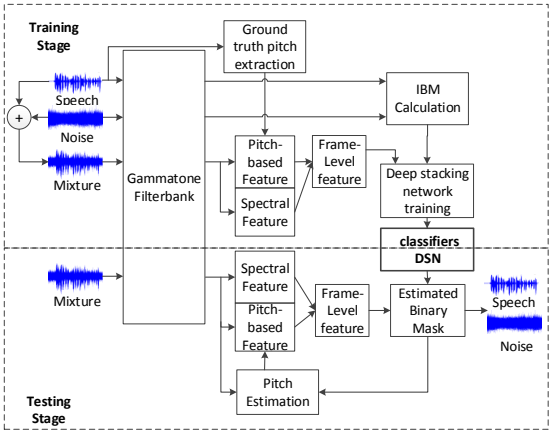


Fig. 1. The architecture of the proposed system.

2.2. Feature Extraction

2.2.1. Signal Decomposition

We first decompose the input signal into a T-F representation. An input mixture sound with 16 kHz sampling rate is decomposed by a 64-channel gammatone filter banks [11] with center frequencies ranging from 50 Hz to 8000 Hz on the equivalent rectangular bandwidth rate scale. The outputs of each channel are divided into 20-ms frame length with a 10-ms frame shift. This processing converts the input signal into a two-dimensional T-F representation, where elements in the representation are called T-F units.

The features used in this work consist of spectral features and pitch-based features extracted from noisy speech.

2.2.2. Spectral features

After decomposing the signal into T-F representation, we compute the energy of T-F unit (c, m) by summing the square of the filter responses in it, and then compress it by a cubic root operation.

$$E(c,m) = \left| \sum_n g^2(c, mT + n) \right|^{1/3}$$

where, c refers to the frequency channel, m refers to the time frame. g is filter responses. And $T = 160$ corresponds to 10 ms frame shift. This energy matrix is called cochleagram.

2.2.3. Pitch-based features

Our pitch-based features are derived from normalized autocorrelation functions (ACF) and envelope ACF. ACF (A) and envelope ACF (A_E) are computed as below.

$$A(c,m,\tau) = \frac{\sum_n g(c,mT-n)g(c,mT-n-\tau)}{\sqrt{\sum_n g(c,mT-n)^2 g(c,mT-n-\tau)^2}}$$

$$A_E(c,m,\tau) = \frac{\sum_n e(c,mT-n)e(c,mT-n-\tau)}{\sqrt{\sum_n e(c,mT-n)^2 e(c,mT-n-\tau)^2}}$$

where, e is the envelope of g. And delay $\tau \in [0, 12.5ms]$. The delay corresponds to pitch period, and the maximum corresponds to 80 Hz pitch. The ACF is called correlogram [11, 12], which is widely used for pitch estimation and source separation. Envelope ACF depicts the amplitude modulation rate in high-frequency channels [12].

Given the pitch period τ_m at frame m, $A(c, m, \tau_m)$ is a quantitative measure of how the observed signal in T-F unit (c, m) is consistent with τ_m. This measure has already been used and proven to be effective under anechoic conditions. To model the high-frequency channel better, we also use $A_E(c, m, \tau_m)$. If there is no target pitch at frame m, both of $A(c, m, \tau_m)$ and $A_E(c, m, \tau_m)$ are set to zero. The pitch-based features we use are $A(c, m, \tau_m)$ and $A_E(c, m, \tau_m)$.

In the training stage, the ground truth pitch is used which is extracted from the clean speech using Praat [13]. And in the test stage, the pitch is extracted by the method described in section 2.4.1.

2.2.4. Frame-level features

To estimate the IBM, we train a classifier using the frame-level features, which are formed by combining the unit-level features of all channels at a frame. The frame-level features include spectral features (E) and pitch-based features $(A,\ A_E)$. In this study, the frame-level features are 192-dimentional consisting of 64-dimension spectral features and 128-dimension pitch-based feature.

2.2.5. Pre-processing

The varieties of noise lead to a very large input feature space, which is hard to model for learning algorithm. To make the learning easier, we use spectral subtraction [14] to restrict the input space as a pre-processing. Although spectral subtraction

makes a stationary assumption about interference, the results of our experiments show that this pre-processing still has some positive effects on non-stationary noise.

2.3. Classifier

We use DSN [15] as the classifier in our system. DSN provides a method to stack simple processing modules for building deep architectures. In DSN, each module is a perceptron followed by a non-linear transformation (seen in Fig. 2(a)). The lower module's output is treated as a part of input to the adjacent higher module, as the illustration in Fig. 2(b).

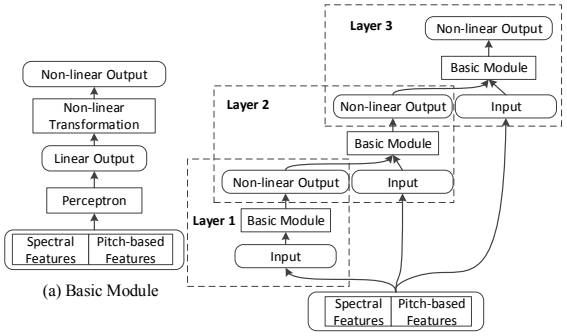


Fig. 2. The architecture of DSN with three layers.

The non-linear transformation is a sigmoid function $\sigma(x) = \frac{1}{1+e^{-(x-\phi)}}$, where ϕ is bias. In this study, we tune the bias to maximize hit minus false alarm rates (HIT-FA) on the training set. HIT-FA [9] is a widely used assessment criteria for speech separation. We use a grid search to find the best bias in the range between -2 and 2 with 0.1 steps.

2.4. DSN for Pitch Estimation & Speech Separation

2.4.1. Pitch estimation with mask

The summary correlogram is computed for pitch estimation, which is calculated as below:

$$S(m,\tau) = \sum_c (A(c,m,\tau) + A_E(c,m,\tau)) \cdot L(c,m)$$

where $L(c,m)$ is 0 or 1 which is the value of the estimated IBM at T-F unit (c,m). The pitch period of the target speech at frame m, τ_m, is the lag corresponding to the maximum of $S(m,\tau)$ in the plausible pitch range [2 ms, 12.5 ms].

Since the estimated IBM also includes unvoiced speech which has no pitch, we distinguish the voiced and unvoiced speech segments with a threshold related to $L(c,m)$. If $S(m,\tau) > \theta \sum_c L(c,m)$, frame m is marked as voiced, else marked as unvoiced. Here $\theta = 0.2$.

2.4.2. Embedding in DSN

DSN can be viewed as a data process pipeline. Since adjacent modules are not tightly combined, we can add some external process to update the input to the next module. After getting an estimated IBM, we re-estimate pitch and update the pitch-based features using the method described in Section 2.4.1. And then the updated features are feed into the next module. For the lowest module, we set pitch to 0 for both training and testing.

3. SYSTEMATIC EVALUATION AND COMPARISON

3.1. Experimental Setup

3.1.1. Dataset

We use clean speech corpus of Chinese National Hi-Tech Project 863 corpus, which consists of 100,000 utterances recorded by 200 speakers. The interference noise includes 16 different types of noise: n1-machine operation, n2-cocktail party noise, n3-factory, n4-siren, n5-speech shaped noise, n6-white noise, n7-bird chirp, n8-crow, n9-crowd, n10-babble, n11-engine start, n12-alarm, n13-playground, n14-traffic, n15-water, n16-wind. These noises cover a variety of daily noises and most of them are highly non-stationary. All signals are down-sampled to 16 kHz. We randomly select 50 utterances from a female speaker and mix them with 6 noises (n1-n6) at 0 dB to set up our training set. And another 200 utterances from the same speaker are mixed with all 16 noises (n1-n16) at -10, -5, 0, 5 and 10 dB as our test set.

3.1.2. Related Methods for Comparison

In order to systematically evaluate the proposed system, we compare it with some other systems: GMM-based [5], DNN-based [10] and DNN-SVM-based [10] methods. For GMM-based method (denoted as 'GMM'), we use a 64-components GMM with diagonal covariance. For DNN-based method (denoted as 'DNN'), we use a DNN with two 200-nodes hidden layers, which is trained by mini-batch gradient descent method with 200 epochs for RBM pre-training and with 100 epochs for network fine-tuning. For DNN-SVM-based method (denoted as 'DNN-SVM'), we combine raw features and the outputs of the last hidden layer in DNN to train a linear SVM. All these 3 methods train a classifier for each channel using unit-level features. The unit-level features include 15-D AMS, 13-D RASTA-PLP, 31-D MFCC and 6-D pitch-based features [5]. The pitch is provided by a multi-pitch tracker [16]. The features are 65-D in total.

The proposed method (denoted as 'Proposed') is a DSN with 5 basic modules. The features we used were frame-level features with a context of 5 frames. In order to measure the effects of our pitch updating procedure, we use ground truth pitch and estimated pitch by [16] to replace our pitch updating procedure. We denote the modified version of the proposed method with truth pitch and estimated pitch as '*Proposed-M-T*' and '*Proposed-M-E*' respectively.

3.2. Experiments

3.2.1. Separation without pitch update

In this subsection, we want to show the ability of different methods in the use of pitch information. First, the ground truth pitch is employed for all systems. The experiment here uses mixed utterances at 0 dB from the test set. The HIT-FA results are shown in Fig. 3. We see that the 'Proposed-M-T' obtains the best HIT-FA results on both matched and unmatched noisy conditions, and difference between two conditions is also the smallest. It means that the proposed system can take the most advantage of a ground truth pitch.

Second, the estimated pitch extracted by a multi-pitch tracker [16] is used in all systems. The results are shown in table 1. From table 1, we can see the similar results when using the estimated pitch. Obviously, the estimated pitch is not better than the ground truth pitch, so the 'Proposed-M-E' results drop down but are still comparable with the 'DNN-SVM' which is the best one except the 'Proposed'. And the 'Proposed-M-E' significantly outperforms the 'DNN-SVM' in unmatched-noise conditions.

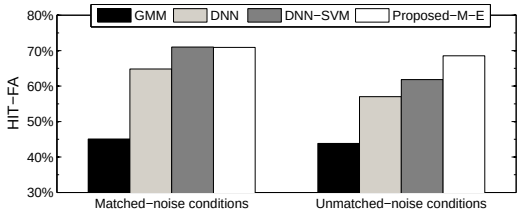


Fig. 3. Overall HIT-FA results of different methods at 0 dB with ground truth pitch on test set.

Table 1. Performances of different methods at 0 dB.

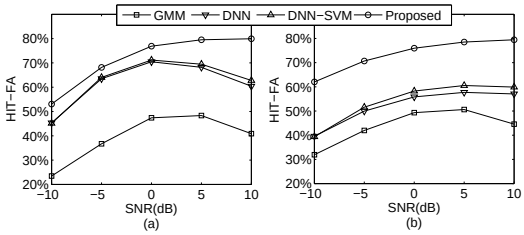
	Methods	HIT	FA	HIT-FA
Matched	GMM	78.88	31.48	47.40
	DNN	**85.64**	15.18	70.45
	DNN-SVM	85.29	14.13	71.16
	Proposed-M-E	84.39	13.47	70.91
	Proposed	85.00	**8.12**	**76.88**
Unmatched	GMM	79.02	29.69	49.33
	DNN	87.62	31.77	55.85
	DNN-SVM	87.17	28.93	58.24
	Proposed-M-E	**87.94**	19.41	68.53
	Proposed	87.81	**11.83**	**75.97**

3.2.2. Separation with pitch update

In this subsection, we want to show the efficiency of our pitch update procedure. We revisit the Table 1. The 'Proposed' denotes our method without any modification. We can see: 1) The 'GMM' is the worst which indicates that deep architecture is likely more suitable for the speech separation problem than a shallow one. 2) The 'DNN-SVM' outperforms the 'DNN', which mainly owes to the generalization ability of SVM. 3) The 'Proposed' remarkably outperforms the other comparison methods on both matched and unmatched conditions. 4) The proposed method has good generalization ability on unmatched conditions. There is only a small gap between matched and unmatched conditions.

3.2.3. Generalization

The experiments described in this subsection use the whole test set includes all 5 different SNR conditions. We examine the generalization of separation systems on unmatched SNR conditions. The HIT-FA results are shown in Fig. 4. From Fig. 4, we can see that the proposed algorithm achieves the best generalization performances at all of the SNR conditions. It also can be observed that the HIT-FA rates of the proposed algorithm become higher with increasing SNR. While other comparison methods achieve the best results at 0 dB (matched SNR condition). This is mainly because we use spectral subtraction as a pre-processing, which restricts the input feature space and make the modeling easier.



(a) Matched-noise condition (b) Unmatched-noise condition

Fig. 4. Overall HIT-FA performances on the different SNR test conditions.

4. RELATED WORK

Some previous researches [5, 10] treated pitch estimation as a preliminary work for separation. In [17], the authors treated the speech separation and pitch extraction as a *"chicken-and-egg"* problem as we do in this study. But we combine these two problems more concisely by using DSN.

5. CONCLUSIONS AND FUTURE WORKS

In this study, we treated speech separation and pitch extraction as a *"chicken-and-egg"* problem. To deal with it, we execute speech separation and pitch extraction alternately. We propose a classification-based algorithm and implement it with DSN. Unlike the conventional separation methods based on DNN directly learning the mapping from input features to target outputs, DSN provides several mid-level separation results. We employ these separation results to refine the pitch estimation. As we known, pitch is very useful information in speech separation. The experimental results show that the proposed algorithm outperforms the conventional methods using DNN or DNN-SVM.

As we seen, our pitch estimation method is relatively simple. Many methods used pitch variance constraint for pitch tracking. In our future work, we can add a tracking process as other pitch estimation algorithm [18]. With pitch tracking, the proposed method is promising to have better performances in speech separation and pitch estimation.

6. REFERENCES

[1] Yang Shao, Soundararajan Srinivasan, Zhaozhang Jin, and DeLiang Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language*, vol. 24, no. 1, pp. 77–93, 2010.

[2] Harvey Dillon, *Hearing aids*, Thieme, 2001.

[3] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[4] Michael L Seltzer, Bhiksha Raj, and Richard M Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[5] Zhaozhang Jin and DeLiang Wang, "A supervised learning approach to monaural segregation of reverberant speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 625–638, 2009.

[6] William Hartmann, Arun Narayanan, Eric Fosler-Lussier, and DeLiang Wang, "A direct masking approach to robust ASR," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 1993–2005, 2013.

[7] Xiaojia Zhao, Yuxuan Wang, and DeLiang Wang, "Robust speaker identification in noisy and reverberant conditions," 2014.

[8] Kun Han and DeLiang Wang, "Towards generalizing classification based speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 168–177, 2013.

[9] Gibak Kim and Philipos C Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2080–2090, 2010.

[10] Kun Han and DeLiang Wang, "An SVM based classification approach to speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4632–4635.

[11] DeLiang Wang and Guy J Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *Neural Networks, IEEE Transactions on*, vol. 10, no. 3, pp. 684–697, 1999.

[12] Guoning Hu and DeLiang Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *Neural Networks, IEEE Transactions on*, vol. 15, no. 5, pp. 1135–1150, 2004.

[13] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[14] S Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79*. IEEE, 1979, vol. 4, pp. 200–203.

[15] Li Deng, Dong Yu, and John Platt, "Scalable stacking and learning for building deep architectures," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2133–2136.

[16] Zhaozhang Jin and DeLiang Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1091–1102, 2011.

[17] Guoning Hu and DeLiang Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2067–2079, 2010.

[18] Sira Gonzalez and Mike Brookes, "Pefac - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.