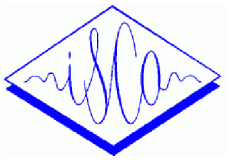
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Boosting Classification Based Speech Separation Using Temporal Dynamics

Yuxuan Wang[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive Science, The Ohio State University, USA
{wangyuxu,dwang}@cse.ohio-state.edu

Abstract

Significant advances in speech separation have been made by formulating it as a classification problem, where the desired output is the ideal binary mask (IBM). Previous work does not explicitly model the correlation between neighboring time-frequency units and standard binary classifiers are used. As one of the most important characteristics of speech signal is its temporal dynamics, the IBM contains highly structured, instead of, random patterns. In this study, we incorporate temporal dynamics into classification by employing structured output learning. In particular, we use linear-chain structured perceptrons to account for the interactions of neighboring labels in time. However, the performance of structured perceptrons largely depends on the linear separability of features. To address this problem, we employ pretrained deep neural networks to automatically learn effective feature functions for structured perceptrons. The experiments show that the proposed system significantly outperforms previous IBM estimation systems.

Index Terms: Monaural speech separation, temporal dynamics, structured perceptron, deep neural networks

1. Introduction

Monaural speech separation is a central problem in speech processing. Since no spatial information can be used, monaural speech separation can only use intrinsic properties of speech or noise, and is thus challenging. In this study, we only consider separating speech from nonspeech interference.

Computational auditory scene analysis (CASA) attempts to solve the speech separation problem based on perceptual principles. It offers several advantages over the traditional speech enhancement methods, e.g., no stationarity is assumed. A primary computational goal of CASA is the estimation of the ideal binary mask (IBM) [9]. The IBM is defined as a mask in which each time-frequency (T-F) unit is labeled as 1 or 0 based on the local signal-to-noise ratio (SNR). If the local SNR of a unit exceeds a threshold, the unit is labeled as 1, otherwise 0. Recent efforts have been made in the CASA community to formulate IBM estimation as a binary classification problem. This formulation has achieved notable success in robust automatic speech recognition [8] and improving human speech intelligibility in noise [6]. Recent work has significantly improved classification performance by exploiting the selection of features [10] and classifiers [3].

Dictated by the speech production mechanism and linguistic constraints, speech signal contains rich temporal information, which could be exploited for speech separation. Systems accounting for temporal dynamics exist. For example, Mysore et al. [7] directly model temporal dynamics using hidden Markov models (HMMs). As a result of temporal continuity[1], the IBM contains highly structured patterns. However, none of the above classification based systems explicitly model temporal dynamics and each T-F unit is labeled without considering neighboring labels. To address this deficiency, we propose to use structured output learning models that are capable of capturing interactions between labels. In particular, we employ linear-chain structured perceptrons [1], which are discriminative Markov random fields trained by the averaged perceptron algorithm. However, structured perceptrons are linear models with limited modeling power. To deal with this limitation, we further employ pretrained deep neural networks to learn highly nonlinear feature functions for structured perceptrons.

In the next section, we describe the proposed system including the system overview, temporal dynamics modeling and nonlinear feature function learning. Experimental results are shown in Section 3. We conclude this paper in Section 4.

2. Proposed Method

2.1. System Overview

A sound mixture with 16 kHz sampling rate is passed through a 64-channel gammatone filterbank with center frequencies ranging from 50 Hz to 8000 Hz. The output from each channel is divided into 20-ms frames with 10-ms frame shift, producing a cochleagram. The computational goal here is to estimate the IBM for the

[1]Other contextual constraints such as common onset and comodulation also contribute to the structure in the IBM, but we only consider temporal dynamics in this study.

mixture. Due to different spectral properties across frequency channels, we train different classifiers for different channels, with the IBM providing training labels. In our previous work [10], we have identified a set of T-F unit level complementary features that are effective for separation. Unit level feature extraction is possible because a T-F unit is a subband signal of a certain length. In this study, we also employ this set of features, which consists of amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), mel-frequency cepstral coefficients (MFCC), pitch-based features and delta features. We use ideal pitch in training but estimated pitch in testing.

2.2. Incorporating Temporal Dynamics

Previous classification based systems mainly use Gaussian mixture models (GMMs) [6] and support vector machines [3, 10] as subband classifiers. However, they do not explicitly model the correlation in time. While delta features can capture some temporal variations on the feature level, each T-F unit is still classified individually. On the other hand, structured output learning models generalize traditional classifiers to predict structured objects such as sequences and trees. In this study, we treat unit labeling at each channel as a sequence labeling problem and employ linear-chain structured perceptrons [1] as the subband classifier. Structured perceptrons generalize the standard perceptron algorithm to predict outputs structured on a random field. Unlike HMM, a structured perceptron is a discriminative model and does not need the independence assumption of features, making it more suitable to our classification task. Compared to conditional random fields, it is more efficient as it does not need to compute the partition function.

For simplicity, in this study we only consider the interaction between a label and its predecessor. Therefore, the discriminant function of a structured perceptron can be written as follows:

$$F(\mathbf{y}, \mathbf{x}) = \sum_i \mathbf{w}^T \phi_1(y_i, \mathbf{x}) + \mathbf{v}^T \phi_2(y_{i-1}, y_i, \mathbf{x}), \quad (1)$$

where i is the frame index, $\mathbf{y}$ and $\mathbf{x}$ are output (label) and input (feature) sequences, respectively. ϕ_1 and ϕ_2 are vector-valued association and interaction feature functions, respectively. Association feature functions define the local discriminant functions for individual T-F units. Interaction feature functions complement association ones by capturing the interactions between neighboring T-F units in time. For example, with interaction feature functions, it is possible to learn when two neighboring T-F units should be assigned to the same label. $\mathbf{w}$ and $\mathbf{v}$ are parameters to be learned by the standard perceptron training algorithm. The predicted sequence labeling is obtained via

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} F(\mathbf{y}, \mathbf{x}).$$

If the predicted sequence labeling $\hat{\mathbf{y}}$ is different from the training sequence labeling $\mathbf{y}^*$ under the current model, i.e., if $\max_{\mathbf{y}} F(\mathbf{y}, \mathbf{x}) > F(\mathbf{y}^*, \mathbf{x})$, then we update the model parameters using the following equations:

$$\begin{aligned} \mathbf{w} &= \mathbf{w} + \sum_i \left[\phi_1(y_i^*, \mathbf{x}) - \phi_1(\hat{y}_i, \mathbf{x})\right], \\ \mathbf{v} &= \mathbf{v} + \sum_i \left[\phi_2(y_{i-1}^*, y_i^*, \mathbf{x}) - \phi_2(\hat{y}_{i-1}, \hat{y}_i, \mathbf{x})\right]. \end{aligned}$$

In our task, we directly use T-F unit feature vectors as the association feature functions and their concatenations as the interaction feature functions:

$$\begin{aligned} \phi_1 =& [\delta(y_i = 0)\mathbf{x}_i, \delta(y_i = 1)\mathbf{x}_i]^T, \\ \phi_2 =& [\delta(y_{i-1} = 0, y_i = 0)\mathbf{z}_i, \delta(y_{i-1} = 0, y_i = 1)\mathbf{z}_i, \\ & \delta(y_{i-1} = 1, y_i = 0)\mathbf{z}_i, \delta(y_{i-1} = 1, y_i = 1)\mathbf{z}_i]^T, \end{aligned}$$

where δ is the indicator function, $\mathbf{x}_i$ is the feature vector of the i-th unit, and $\mathbf{z}_i = [\mathbf{x}_{i-1}, \mathbf{x}_i]^T$.

As can be seen, sequence decoding is involved in both training and testing. In our setting, we could simply use the Viterbi algorithm for efficient decoding. For other complex interaction forms, more sophisticated decoders can be used, but with significantly higher complexity.

2.3. Learning Nonlinear Feature Functions

Structured perceptrons are linear models with limited modeling power. The performance is largely dependent on the linear separability of features. Unfortunately, acoustic features are usually not linearly separable. To address this issue, we further propose to use pretrained deep neural networks (DNNs) to automatically learn feature functions that can greatly increase the modeling power of structured perceptrons.

Deep neural networks have received widespread attention since Hinton et al.'s 2006 paper [5]. Deep neural networks can be viewed as hierarchical feature detectors that learn increasingly complex feature mappings as the number of hidden layers increases. However, training deep neural networks with backpropagation is difficult due to problems such as vanishing gradients. To alleviative this problem, Hinton et al. propose to first pretrain a DNN using a stack of unsupervised, restricted Boltzmann machines (RBMs) in a layerwise fashion before performing backpropagation on any objective function of interest. Raw features are used to train the first RBM, whose hidden activations are then treated as the new training data for the second RBM, and so on. The resulting network weights are used to initialize a DNN with the same depth and size. It has been shown that such a generative-discriminative process is crucial for successfully training

a deep architecture. DNNs have achieved state-of-the-art performance on many pattern recognition tasks, including automatic speech recognition (e.g. [2]).

An RBM is a two layer neural network, with connections only between its visible layer $\boldsymbol{v}$ and hidden layer $\boldsymbol{h}$. It has an energy function E defining joint probability $p(\boldsymbol{v}, \boldsymbol{h}) = e^{-E(\boldsymbol{v},\boldsymbol{h})}/Z$, where $\boldsymbol{v}$ and $\boldsymbol{h}$ denote a visible and hidden layer configuration, respectively. Z is the partition function to ensure that $p(\boldsymbol{v}, \boldsymbol{h})$ is a valid probability distribution. We use a Gaussian-Bernoulli (i.e. Gaussian visible layer and Bernoulli hidden layer) RBM for the first layer of a DNN and Bernoulli-Bernoulli RBMs for all the layers above. The energy function of a Gaussian-Bernoulli RBM is defined as follows and the interested reader is referred to [5] for the Bernoulli-Bernoulli case,

$$E(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{2}\sum_i (v_i - a_i)^2 - \sum_j b_j h_j - \sum_{i,j} w_{ij} v_i h_j,$$

where v_i and h_j are the ith and jth units of $\boldsymbol{v}$ and $\boldsymbol{h}$, a_i and b_i are the biases for v_i and h_j, respectively, and w_{ij} is the symmetric weight between h_j and v_i.

Maximum likelihood estimation can be used to train an RBM. Similar to the standard results for log-linear models, the gradient of the log-likelihood is the difference between the expectation under the empirical distribution and the expectation under the model distribution. The calculation of the second expectation involves exponentially many terms, but it can be effectively approximated using contrastive divergence [4].

After RBM pretraining, we fine-tune the whole network using backpropagation with a cross-entropy loss function to make the network discriminative. Since the last layer of such a network essentially defines a linear classifier, the last hidden layer representation is likely more linearly separable if the network is well trained. Therefore, we can take the hidden activations from the last hidden layer as the automatically learned feature function for the structured perceptron. In other words, the discriminant function of a structured perceptron now becomes:

$$F_g(\mathbf{y}, \mathbf{x}) = \sum_i \mathbf{w}^T \phi_1(y_i, g(\mathbf{x})) + \mathbf{v}^T \phi_2(y_{i-1}, y_i, g(\mathbf{x})). \quad (2)$$

Here a pretrained DNN acts as the discriminative nonlinear mapping $g(\cdot)$. This way structured perceptrons would greatly benefit from the nonlinear modeling power of deep architectures, which we will show next.

3. Experimental Results

We employ the IEEE corpus recorded by a female speaker for systematic evaluations. For training, we mix 50 sentences with 12 nonspeech noises at 0 dB. The 12 noises are babble, bird chirp, crow, cocktail party, yelling, crowd clap, rain, rock music, siren, telephone, white, and wind noise. For testing, we use 20 new sentences to mix with the 12 noises at 0 dB. To test generalization of the proposed system, we also create an unmatched test set by mixing the 20 test sentences with 3 unseen noises: speech-shaped, traffic, and factory noise. The test noises cover a variety of daily noises and most of them are highly nonstationary.

We compare with two previous classification based speech separation systems [6, 3] to demonstrate the effectiveness of the proposed system (named DNN-Struct). We also present results from two baseline systems that are based on DNNs or structured perceptrons (named Struct-Perc) alone, in order to disentangle the contribution of each component in the system. We use two hidden layer DNNs and fine-tune the whole network using the limited-memory BFGS algorithm (L-BFGS) after 100 epochs of RBM pretraining. Structured perceptrons are trained for 50 epochs and the final models use averaged parameters. We employ classification accuracy as well as hit minus false-alarm (HIT$-$FA) rate as the evaluation criteria in this study. Here, the HIT rate is the percent of correctly classified target-dominant T-F units (1s) in the IBM. The FA rate is the percent of wrongly classified interference-dominant (0s) T-F units in the IBM. The HIT$-$FA rate is proposed in [6] and shown to be highly correlated with human speech intelligibility.

We report HIT$-$FA results at three kinds of frames: overall, voiced and unvoiced. Voicing boundaries are determined based on ideal pitch of speech. Table 1 shows the classification performance of different systems on the matched-noise test set. First, it is instructive to directly compare the classification performance between SVM [3] and DNN. We use the same feature set to train both SVM and DNN, and clearly DNN outperforms SVM in terms of both accuracy and HIT$-$FA. The performance improvement is particularly large for unvoiced speech, which is harder to separate due to the lack of harmonics and weak energy. This result suggests that deep architectures are likely more suitable for the speech separation problem than shallow ones. We note that DNNs without RBM pretraining (i.e., standard multi-layer perceptrons) produce significantly worse results in our experiments. Structured perceptrons are able to model temporal dynamics, but only with linear modeling capability. As can be seen in Table 1, the performance is actually significantly worse than standard binary classifiers such as SVM. Nevertheless, the performance is substantially boosted by using learned nonlinear feature functions with DNNs. The proposed system significantly outperforms other comparisons in terms of both accuracy and HIT$-$FA, and the improvement over Kim et al.'s GMM based system is quite large. Kim et al.'s system has been shown to improve speech intelligibility in noise [6], it is therefore reasonable to project that the proposed

Table 1: Classification performance of different systems on a matched-noise test set. Boldface indicates best result

System	Overall			Voiced			Unvoiced			Accuracy
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA	
Kim et al. [6]	75.4%	20.0%	55.4%	79.1%	22.0%	57.1%	57.4%	16.4%	41.0%	77.4%
SVM [3]	75.7%	7.7%	68.0%	80.5%	7.5%	73.0%	56.7%	8.2%	48.5%	86.6%
Struct-Perc	72.3%	12.5%	59.8%	77.9%	11.5%	66.4%	50.0%	14.2%	35.8%	82.3%
DNN	79.0%	7.5%	71.5%	82.4%	8.0%	74.4%	68.2%	8.1%	60.1%	87.5%
DNN-Struct	82.1%	7.0%	**75.1%**	84.3%	7.1%	**77.2%**	72.0%	7.0%	**65.0%**	**89.1%**

Table 2: Classification performance of different systems on an unmatched-noise test set

System	Overall			Voiced			Unvoiced			Accuracy
	HIT	FA	HIT−FA	HIT	FA	HIT−FA	HIT	FA	HIT−FA	
Kim et al. [6]	65.7%	34.6%	31.1%	67.4%	34.9%	32.5%	53.6%	34.0%	19.6%	66.1%
SVM [3]	64.8%	4.0%	60.8%	68.8%	4.7%	64.1%	41.0%	2.7%	38.4%	90.9%
Struct-Perc	65.7%	4.7%	61.0%	71.5%	5.7%	65.8%	31.3%	2.7%	28.6%	90.6%
DNN	68.6%	5.2%	63.4%	70.2%	5.8%	64.4%	59.0%	4.0%	55.0%	90.6%
DNN-Struct	71.0%	4.6%	**66.4%**	71.8%	5.1%	**66.7%**	65.8%	3.7%	**62.1%**	**91.4%**

Table 3: SNR and SegSNR results of different systems

Criteria	SNR (dB)		SegSNR (dB)	
	Matched	Unmatched	Matched	Unmatched
Kim et al. [6]	10.2	6.8	7.3	2.2
SVM [3]	10.5	8.8	10.9	7.2
DNN-Struct	**12.2**	**9.5**	**13.2**	**8.2**

system will provide further improvements.

Due to the mismatch between training and testing, classification becomes harder on the unmatched-noise test set. As we can see in Table 2, Kim et al.'s system fails to generalize due to substantially increased FA rates. The proposed system consistently outperforms the others and achieves reasonable performance on unseen noises. The improvement of DNN-Struct over DNN indicates that modeling temporal dynamics improves the generalization of the classifiers.

In addition to classification performance, Table 3 also presents SNR and segmental SNR (SegSNR) comparisons, which are standard criteria for speech enhancement algorithms. The target speech resynthesized from the IBM is used as the ground truth signal. By the SNR based criteria the proposed system also significantly outperforms previous systems. Moreover, the proposed system produces more natural sounding output due to the smoothing effect from temporal dynamics modeling.

4. Conclusions

We have proposed a classification based speech separation system that explicitly models temporal dynamics. We formulate the separation problem as a sequence labeling problem and employ structured averaged perceptrons. We transform the standard structured perceptron into a highly nonlinear sequence classifier by using feature functions learned from pretrained DNNs. The proposed system significantly outperforms previous systems in terms of classification accuracy, HIT−FA, and SNR. To our knowledge, this is the first study that uses DNNs to address the speech separation problem. In the current study, we only model the interaction between two neighboring frames. In future work, we will investigate more complex forms of temporal interaction, including long-range temporal interactions.

Acknowledgements. This research was supported in part by an AFOSR grant (FA9550-08-1-0155) and an STTR grant from AFOSR.

5. References

[1] M. Collins, "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *Proc. EMNLP*, 2002.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE T-ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[3] K. Han and D. Wang, "An SVM based classification approach to speech separation," in *Proc. ICASSP*, 2011, pp. 5212–5215.

[4] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comp.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[5] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comp.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[6] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *JASA*, vol. 126, pp. 1486–1494, 2009.

[7] G. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. ICASSP*, 2011, pp. 17–20.

[8] M. Seltzer, B. Raj, and R. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Comm.*, vol. 43, no. 4, pp. 379–393, 2004.

[9] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Divenyi P., Ed. Kluwer Academic, Norwell MA., 2005, pp. 181–197.

[10] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech separation," Ohio State Univ. Dept. of CSE, Tech. Rep. TR37, 2011.