Speech Separation from Noises by using Deep Neural Network

P.Nancy[1]
Post graduated.,Dept of ECE
Parisutham institute of technology and science,thanjavur.
Affiliated to Anna university,Chennai,India.
Email: nancyece3@gmail.com

D. Sellathambi[2]
Assistant professor.,Dept of ECE
Parisutham institute of technology and science,thanjavur.
Affiliated to Anna university,Chennai,India.
Email: sellathambil@gmail.com

***Abstract*-Speech separation from noise is a binary classification problem. To solve this problem the speech signal processed with variety of matching and un-matching conditions. To train the various training data set kernel support vector machine was used. The large variety of training data sets processed by using deep neural network is proposed here. The training data set processed and compared with various other data sets to estimate the pitch of the signal based on Ideal binary mask (IBM) construction. Various type of filter is used for process the signal. The signal processed at the various hidden layers. The signal processed at the different types of frequency ranges. A different type of classifier is used for process the signal based on the difference between the HIT –FA (influence of the hit-false alarm) rate is used for analyze the signal. In existing system Back propagation algorithm is used for propagate the signal at 'n' input layer. It perform mapping between the input and output. This algorithm provides better output with less error. In proposed method Semi-supervised learning algorithm is used to improve speech intelligibility in background noise by providing more gain for soft low frequency sounds than for loud low frequency sounds. Simple implementation and Computation time will be reduced. It provides support for the unknown noise also and it separates the noise about 80% or even more. The speech separation system mainly used in automatic speech recognition and high hearing aids. In kernel support vector machine, it support for only known noise. But in our proposed system, it support unknown noise also.**

Keyword: binary classification problem, IBM, HIT-FA rate, semi-supervised learning algorithm

I. Introduction

Speech separation has used in variety of application such as hearing aids design, robust automatic speech recognition (ASK) and mobiles phones. However, separating the speech from general acoustic environments is a big challenge. Monaural speech Separation is particularly difficult to separate the speech from noisy signal. Monaural speech separation is a Combination of both high frequency and low frequency with different harmonic. In this case, particularly speech signal separation is very difficult. In this paper, we focused on monaural speech separation from non-speech background interference. Spectral subtraction [5] is a classical method for noise reduction, which subtract an estimate the noise spectrum from different mixture spectrum. Computational Auditory Scene Analysis (CASA) will separate the sound mixture into different auditory streams especially dealing with unvoiced harmonic structure. CASA system has limited capability the ideal binary mask (IBM) is a time-frequency(T-F)mask constructed from pre-mixed speech and noise[4] this mask define in terms of premixed target dominant and interference dominant specifically for time frequency unit. If the signal to noise ratio within the unit is greater than local SNR criterion (LC) we call it target dominant and corresponding mask element in the IBM is set to '1'.otherwise the mask element is set to '0' and we call the unit interference dominant, IBM is defined as;

$$IBM(t,f) = \begin{cases} 1, if\ SNR\ (t,f) > LC \\ 0, \qquad otherwise \end{cases} \qquad (1)$$

Where, SNR(t,f) denotes the local SNR(in decibels) with in the T-F unit at time 't' and frequency 'f' the effectiveness of IBM estimation has also been demonstrated for robust ASK [8],[3]. Our task is to estimate the IBM through binary classification. Different speakers, background noises, room reverberation and channel distortions can all introduce severe mismatches between training and test conditions variety of acoustic conditions into the training sets due to the expensive quadratic programming. It support only known noise but it cannot support unknown noise.

The objective of this paper to training with a large variety of acoustic conditions coupled with the use of linear SVMs [9].Deep neural network supports the unmatched test conditions feature learning neural network are pre-trained using restricted Boltzmann machines (RBMs), which generative models and serve as pre-training for the recently proposed deep neural belief networks (DBNs)[4],[5].
Neural network with many hidden layers can be viewed as hierarchical feature detectors that capture higher-order correlations between raw and feature. Training deep neural network using the back propagation algorithm is relatively simple implementation. This algorithm used for propagate the signal for N number of input layers. DBNs pre-train each layer generatively using RBMs this DBMs is effective method and there is an increasing number of successful applications of DBNs first in visual processing [6] and more recently in speech processing [2],[7]. To separate the speech from noise

diagram as shown in Figure 1 separating our original speech from different kind of noises by using deep neural network. Input signal consist of noise and information

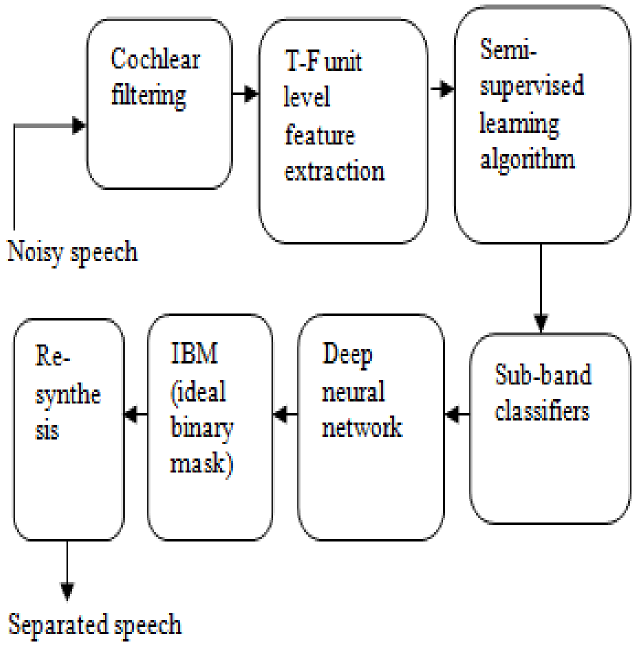


Fig .1: schematic diagram of speech separation system.

Signal these noises are environment noise like traffic, air, cocktail party because of this noises original information signals are collude. Cochlear filter mainly used in human voice purpose of this filter used to determine which sound has to be mask and which one is too audible. This type of filter is the front end of deep neural network, cochlear filter banks were identifier that are commonly used in the speech and audio communities.

Deep neural network consist of number of layers first layer is visible, other numbers of layer not visible because these are hidden layers. Sub-band classifier is used to classify the signal at various acoustic condition this classifier is used to find the HIT-FA at various matching and un-matching condition used for identification purpose.IBM is a time-frequency mask is used for process the signal at various acoustic condition this mask created based on signal to noise ratio at local criterion.re-synthesis processes used for separate the pure signal it will removes the some unwanted signal present at end of the process finally we get the original speech signal.

II. ESTIMATION OF IBM

We aim to estimate the IBM via binary classification. Time frequency unit representation called cochlea-gram [11], which consists of a matrix of T-F units. To estimate the IBM we classify each T-F unit in the cochlea-gram as either target-dominant or interference-dominant through supervised training. The feature set consist of amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), meal-frequency spectral co-efficient and pitch-based feature. RASTA-PLP and pitch-based feature are important for generalization to unseen condition. HIF-FA (influence of the hit-false alarm) is used to estimated the IBM.HIF-FA rate, is the percent of correctly classified target dominant T-F unit in the IBM(set as one).FA rate is percent of wrongly classified interference dominant T-F unit in the IBM(set as zero).we use HIT-FA as our main evaluation criterion for assessing classification-based speech separation systems.

TABLE I. HIT-FA RESULT FOR TWO CLASSIFIERS TRAINED ON DIFFERENT NUMBERS OF NOISES.

Classifier	**Matched condition**			**Unmatched-noise condition**		
	HIT	***FA***	***HIT-FA***	***HIT***	***FA***	***HIF-FA***
S50N3	85.0%	7.4%	77.6 %	82.6%	20.8%	61.8%
S50N12	81.6%	7.4%	74.2 %	78.3%	11.6%	66.7%

.

We train two Gaussian-kernel SVMs on 50 IEEE female utterances mixed with first 3 noises (N1-N3) and then 12 noises (including N1-N3) at 0 db. These two classifiers, which we call S50N3 and S50N12 respectively, are tested in two test condition. Ten new IEEE female utterances (same speaker) are mixed with N1-N3 to create a matched-noise test condition, and 5 unseen noises to create an unmatched-noise test condition, all at 0 dB.

Table I, presents the overall HIT−FA rates for the two classifiers. S50N3 outperforms S50N12 in the matched-noise condition due to higher HIT rates, because it is exclusively trained on N1-N3. However, S50N12 significantly outperforms S50N3 in the un-matched noise test condition due to much lower FA rates.

Next, we examine the situation when the test speaker differs from the training one. We train three classifiers for comparisons. The first and second are trained on the IEEE female and male utterances respectively, while the third is trained on both. Five noises are randomly chosen to mix with the training utterances at 0 dB to create the training set.

The test and training noises are the same but the mixtures of both genders are tested by the three classifiers. Figure 2 shows the HIT−FA rates. We can see that while the first two classifiers perform well matched-speaker performance significantly degrades when tested on a new speaker.

Different speakers, especially different genders, may have different energy distributions across frequency channels, hence posing difficulties for classifiers that are insufficiently trained. In contrast, the behaviour of the third classifier suggests the effectiveness of training on multiple speakers.

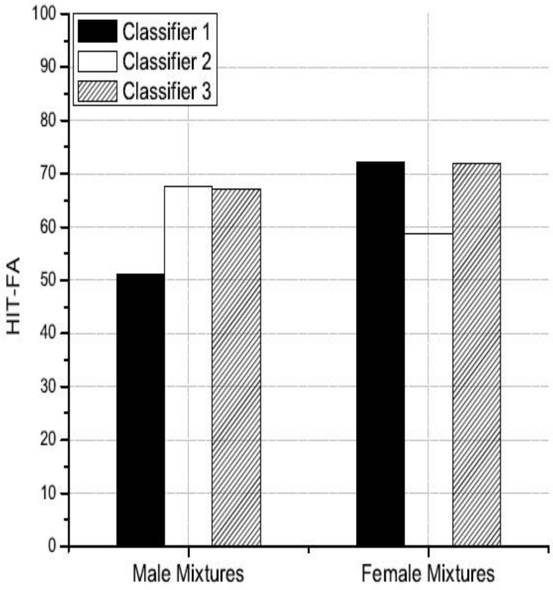


Fig. 2: HIT-FA result tested on different speaker.

The first and second classifiers are trained on the IEEE female and male utterances, respectively the third classifier is trained on both. Sub-band classifiers have both good performance and scalability. Different speakers, especially different genders, may have different energy distributions across frequency channels, hence posing difficulties for classifiers that are insufficiently trained. In contrast, the behaviour of the third classifier suggests the effectiveness of training on multiple speakers.

III. DNN-SVM SYSTEM FOR SPEECH SEPARATION

A. Restricted Boltzmann machines

A restricted Boltzmann machine (RBM) is a generative stochastic neural network that can learn a probability distribution over set of input. RBM are two layer neural networks with a visible layer and a hidden layer[6].RBM have input units, corresponding to feature of their input, hidden units that are trained and each connection in the an RBM must be connected a visible unit to an hidden unit. An RBM has an energy function defining joint probability:

$$p(v,h) = \frac{e^{-E(v,h)}}{z} \qquad (2)$$

Where, v and h denote a visible and hidden layer Configuration; Z is called the partition function to ensure p (v, h) is a valid probability distribution. The hidden layer is binary and hidden units are Bernoulli random variables.But the visible layer v can be either binary or real-valued, the latter being more suitable for modelling acoustic features

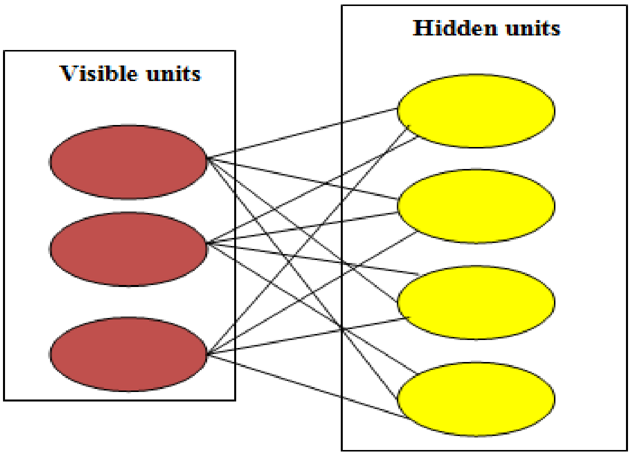


Fig.3: Diagram of a restricted Boltzmann machine with three visible units and four hidden unit (no bias unit).

B. DNN-SVM Architecture

The architecture of the proposed DNN-SVM speech separation system fig [3] DNN-SVM serves as the sub-band classifier. Raw acoustic features are used as training data to train the first RBM, whose hidden activations are then treated as the new training data for the second RBM, and so on. The advantage of RBM pre-training remains even when a large number of training samples are used [10], and it is often critical for training a deep network having many hidden layers.

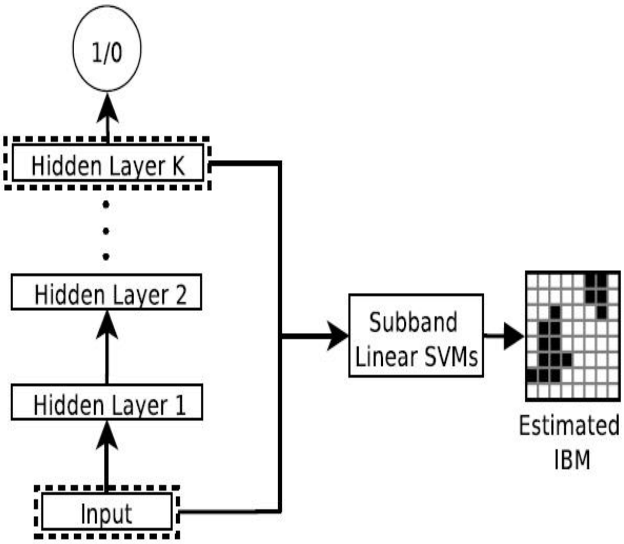


Fig.4: schematic diagram of the proposed DNN-SVM

To make internal representations discriminative, the whole network is then supervised fined-tuned using the back propagation algorithm. We choose the last hidden

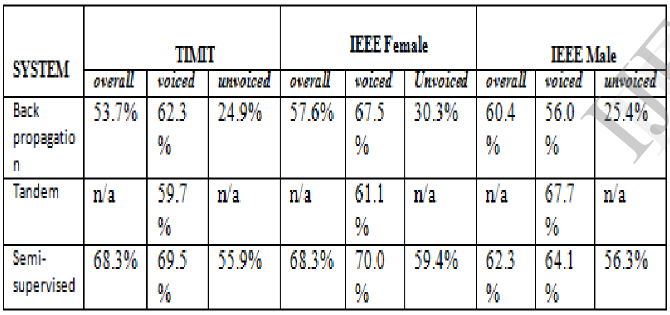
IV. PILOT EXPERIMENTS ON DNN-SVM

DNN-SVM trained on variety of acoustic conditions. Nonlinear feature extraction of DNN is extremely fast. A number of design choices have to be made before training DNN-SVM on large datasets. Here, we present some pilot studies for DNN-SVM using a relatively small corpus, created by mixing 50 IEEE female utterances with 12 randomly chosen noises at 0 dB. The test set is created by mixing 10 new utterances with 12 seen noises (matched-noise test condition) and 10 unseen noises (unmatched-noise test condition) at 0 dB. The training set consists of about 150,000 samples for each channel.

V. RESULTS

The DNN-SVM system performs significantly better than linear SVMs that are trained using the COMB feature set, indicating that discriminatively learning more linearly separable features is indeed needed. Table II reports the HIT−FA rates on 0 dB mixtures. The DNN-SVM system also outperforms the tandem algorithm for voiced speech separation even with ideal sequential grouping, and is much better than with actual sequential grouping. The DNN-SVM system is trained on TIMIT utterances, does not seem to be a problem as demonstrated by the results on the IEEE corpus.

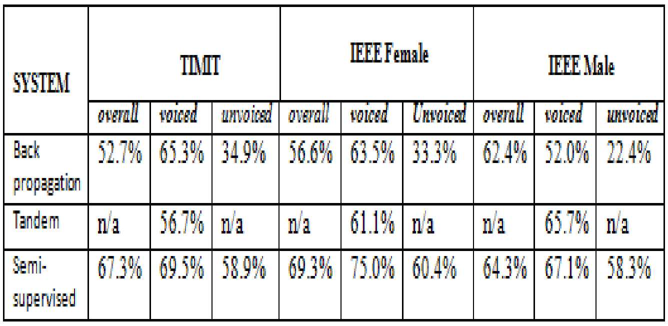
TABLE II: HIT-FA Results of the 0 db TIMIT and IEEE test set.



SYSTEM	TIMIT			IEEE Female			IEEE Male		
	overall	*voiced*	*unvoiced*	*overall*	*voiced*	*Unvoiced*	*overall*	*voiced*	*unvoiced*
Back propagation	53.7%	62.3 %	24.9%	57.6%	67.5 %	30.3%	60.4 %	56.0 %	25.4%
Tandem	n/a	59.7 %	n/a	n/a	61.1 %	n/a	n/a	67.7 %	n/a
Semi-supervised	68.3%	69.5 %	55.9%	68.3%	70.0 %	59.4%	62.3 %	64.1 %	56.3%

Layer activations as the learned features after the network is sufficiently fine tuned. The weights from the last hidden layer to the output layer would essentially define a linear classifies the last hidden layer.

TABLE III: HIT-FA Result on the -5 db TIMIT and IEEE test set



SYSTEM	TIMIT			IEEE Female			IEEE Male		
	overall	*voiced*	*unvoiced*	*overall*	*voiced*	*Unvoiced*	*overall*	*voiced*	*unvoiced*
Back propagation	52.7%	65.3%	34.9%	56.6%	63.5%	33.3%	62.4%	52.0%	22.4%
Tandem	n/a	56.7%	n/a	n/a	61.1%	n/a	n/a	65.7%	n/a
Semi-supervised	67.3%	69.5%	58.9%	69.3%	75.0%	60.4%	64.3%	67.1%	58.3%

We have also used the trained models to estimate the IBM for -5 dB mixtures.HIT−FA rates are reported in Table III. As expected, the results are worse than in Table III but the degradation is not severe. We expect improved results if the systems are also trained on -5 dB mixtures. It would be interesting to see HIT−FA performance as a function of the number of training noises and utterances. We have also used the trained models to estimate the IBM for -5 db mixtures. It would be interesting to see HIT-FA performance as a function of the numbers of training noises and utterances

VI. CONCLUSION

The mismatch problem could be significantly alleviated by training on more acoustic conditions. However, the resulting large training set poses a big challenge to conventional kernel SVMs, which have huge complexity and poor scalability. We have proposed to learn more linearly separable features from raw acoustic features. Linear SVMs are then trained on the combination of learned and raw features to estimate the IBM. We choose deep neural networks for feature learning due to their scalability and flexibility.

REFERENCES

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Audio, Speech, Lang. Process.*,vol. 27, no. 2, pp. 113–120, 1979

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, 2012

[3] W. Hartmann and E. Fosler-Lussier, “Investigations into the incorporation of the ideal binary mask in ASR,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4804–4807

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

[5] *G. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786,p. 504, 2006*

[6] *H. Lee, R. Grosse, R. Ranganath, and A. Ng, “Convolutional deep belief networks for scalable unsupervised learning ofhierarchical representations,” in Proc. the 26th International Conference on Machine Learning, 2009, pp. 609–616*

[7] *A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. Audio, Speech, Lang.Process., vol. 20, no. 1, pp. 14–21, 2012*

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

[9] *S. Shalev-Shwartz, Y. Singer, and N. Srebro, “Pegasos: primal estimated sub-gradient solver for SVM,” in Proc. the 24thInternational Conference on Machine learning, 2007, pp. 807–814*

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

[11] *D. Wang and G. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms and Applications. Hoboken,NJ: Wiley-IEEE Press, 2006*

[12] G. Hu, “100 nonspeech environmental sounds (http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html),” 2004.