ATLANTIS PRESS
Advances in Engineering Research (AER), volume 137
2018 Joint International Advanced Engineering and Technology Research Conference (JIAET 2018)

On the Separability of Local SNR Values Represented by Amplitude Modulation Spectrogram Vectors

Wei Wei, Jinhuan Wang, Jing Wang
School of Intelligence Science and Information Engineering,
Xi'an Peihua University, China
seamus_wei@sina.com, 360635476@qq.com, superj2003@163.com

Keywords: Amplitude modulation spectrograms; Signal to noise ratio (SNR); Deep belief networks; t-distributed stochastic neighbor embedding (t-SNE).

Abstract. Evaluating local SNR values in noisy speech is very valuable for many applications. Research in this paper investigates the separability of local SNR values which are represented by amplitude modulation spectrogram (AMS) feature vectors. Deep belief networks (DBN) were employed in previous studies for binary and quarterly classifications of these SNR feature vectors, however with unsatisfactory results. It is difficult to identify which factor, AMS feature vector or DBN configuration should be responsible for the result. Therefore the technique of t-distributed stochastic neighbor embedding (t-SNE) is called in this study to visualize the separability of AMS feature vectors. According to experimental observations with binary and quarterly classifications, AMS feature vector is a good representation of SNR values with fine grain separability. To improve the classification performance of SNR values represented by AMS feature vectors, more attention should be put on DBN's training and configuration.

1 Introduction

Noise suppression or speech enhancement is an important issue in a wide range of speech processing applications. For example, in the field of automatic speech recognition, background noise is a major problem which typically causes severe degradation of the recognition performance. Strong noise also increases human listener's word error rates in communications.

To reduce noise effect in a variety of applications, many methods have been developed to cancel or decrease noise in noisy speech. We proposed a technique which selects good spectral components from the temporal-frequency representation of a noisy speech in [1]. The basic idea underlying the method is that rebuilding a speech with its partial components which have higher signal to noise ratio (SNR) would lead to better perceptual effect. The method performs well in lower SNR environments, especially with alleviation of music noise which results from most spectral subtraction methods.

In order to evaluate SNR values better in each time-frequency units (frame), deep learning method was introduced in our study in [2]. The initial motivation was a binary classification of all those local SNR values in a noisy speech. These local SNR values correspond to speech frames one by one. For each frame, it is either labeled by its SNR value as good one retained for speech reconstruction or as bad one discarded. Kim et al studied same problems in [3] for intelligibility improvement. The binary classification was performed with a Bayesian classifier. The SNR value of each frame was represented by an amplitude modulation spectrogram (AMS) feature vector. This is a concept introduced by Tchorz et al in [4] for SNR estimation with neural networks. We are very interested in the evaluation scheme in [4] except its shallow neural network. In a trial of replacing shallow neural network with a deep one, as suggested by Hinton in [5], deep belief network (DBN) was introduced in [2]. However, as shown in the paper the result was not satisfactory.

It is difficult to judge which factor, AMS feature vector, or DBN configuration should be responsible for the unsatisfactory result. Therefore the technique of t-distributed stochastic neighbor embedding (t-SNE) is called into the study to demonstrate the separability of AMS feature vectors for subjective judgement. This technique was proposed in [6] by Maaten and Hinton to visualize high

dimensional data. If AMS feature vector can be shown separable in two-dimensional plane by means of t-SNE, it is reasonable to clear AMS features from the source of unsatisfactory result.

The outline of the paper is as follows. In section 2, extractions of AMS feature vectors and the relevant SNR values for time-frequency units in noisy speech are introduced. In section 3, principal of t-SNE used for high dimensional data visualization is explained. Considerations of using t-SNE to show the separability of AMS features and the results are discussed in section 4. Conclusions and suggestions for future work are presented in section 5.

2 Extraction of AMS Feature Vectors and relevant SNR value

According to [4] and [5], AMS feature vectors and relevant SNR values are extracted from speech subband signals frame by frame. Both noisy speech and clean speech are needed for training and testing samples, while in practice only noisy speech is needed for AMS feature extraction. Fig. 1 is the block diagram for AMS feature vectors and local SNR values extraction.

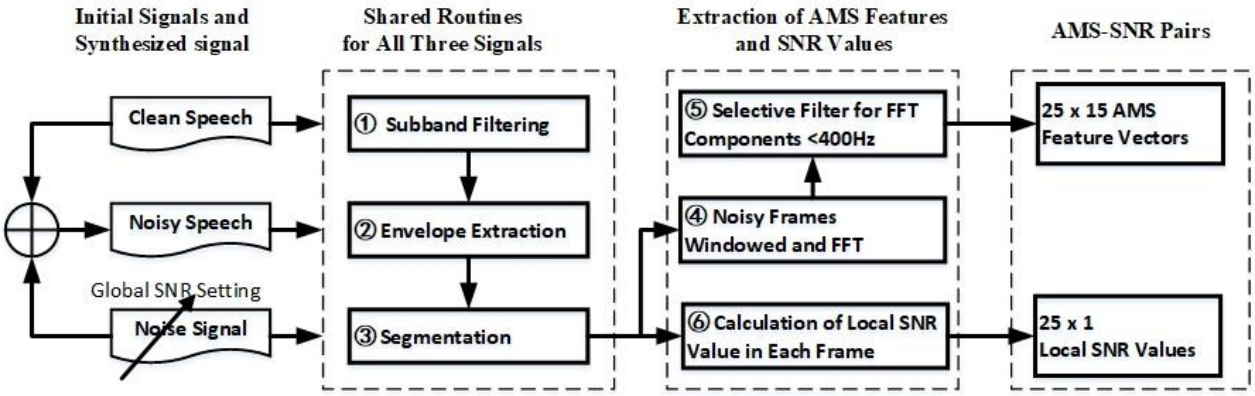

Figure 1. Block diagram for AMS-SNR pair extraction

Initial signals include clean speech and noise signal. The preparation routine is as follows: setting a global SNR value, e.g., -5dB or 0dB, then adjust noise signal to appropriate level in accordance with the assigned global SNR. Noisy speech is created by summing up clean speech and noise signal. The noisy speech and clean speech are used throughout the study.

As shown in Fig.1, AMS feature vectors and the relevant local SNR values are extracted by following steps:

(1) Subband filtering. All three signals, noisy speech, clean speech and noise signal, are filtered by a filter bank and each signal is transformed into 25 subband signals. Henceforth there are 25x3 signals which have equal length.
(2) Envelope Extraction. Down sampling the 25x3 subband signals to get 25x3 envelope signals. The down sampling rate is 2 for sampling frequency at 8kHz, and 3 for 12kHz.
(3) Segmentation. All 25x3 envelope signals are divided into 32ms frames. This is a localization process for the AMS-SNR pairs.
(4) FFT for noisy frames. Frames corresponding to noisy envelope signals are windowed and run FFT for spectral analysis.
(5) Components selection. All frequency components resulted from FFT are chosen by a triangular filter. The filter picks up 15 components whose relevant frequencies are below 400Hz in each subband. This is exactly the 15 dimensional AMS feature vector.
(6) Calculation of local SNR value in each subband frame. The formula is:

$$LSNR_{t,k} = 10\log_{10}\left(\sum_{m=1}^{M} c_{t,k}^2(m) \Big/ \sum_{m=1}^{M} n_{t,k}^2(m)\right) \quad (1)$$

Where $c_{t,k}(\mathrm{m})$ is the amplitude of envelope sample m in clean frame t at sub-band k, $n_{t,k}(\mathrm{m})$ is its counterpart in noise frame. M is the number of envelope samples in each frame.

With AMS feature vectors and the relevant SNR values, there are two choices for DBN training. The first one is to combine AMS vectors from different subbands as one vector and uses it as DBN's

input vector, something similar to what Tchorz did in [4]. The output of DBN is a code of SNR values. According to our observations in the study this is too complex in practice. Therefore, we chose the second way to train DBN. Each AMS-SNR pair is considered as an independent input-output sample, and all 25 subband signals are dealt with separately. Despite of the wide span (-60dB to 60dB in the study) of the detailed SNR values, we are interested only in a binary or a quartering classification of them. Original AMS-SNR pairs can be transformed easily with a threshold or three thresholds for this.

3 High Dimensional Data Visualization by t-SNE

Visualization of high dimensional data is helpful to observe characteristics of the data. It is a natural association that similar data points are close to each other if they are represented in appropriate form. In order to be sure that AMS feature vector is qualified for its role, t-SNE tool [6] is introduced in the study to evaluate the separability of AMS feature vectors subjectively. The t-SNE technique visualizes high dimensional data by giving each data point in high dimensionality a location in a two or three dimensional map. It is a variation of SNE [7] which is much easier to optimize, and produces better visualizations by reducing the tendency to crowd points together in the center of the map.

Similarity is the kernel of SNE. The similarity of a data point x_j to a data point x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i. For high dimensional data, the conditional probability $p_{j|i}$ is given by

$$p_{j|i} = \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right) \Big/ \sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right) \tag{2}$$

Where σ_i is the variance of the Gaussian that is centered on data point x_i. For the low dimensional counterparts y_i and y_j of the high dimensional data points x_i and x_j, the similarity $q_{j|i}$ is given by

$$q_{j|i} = \exp\left(-\left\|y_i - y_j\right\|^2\right) \Big/ \sum_{k \neq i} \exp\left(-\left\|y_i - y_k\right\|^2\right) \tag{3}$$

SNE aims to find a low-dimensional data representation that minimizes the mismatch between $p_{j|i}$ and $q_{j|i}$. It minimizes the sum of Kullback-Leibler divergences over all data points using a gradient descent method. The cost function is given by

$$C = \sum_i KL\left(P_i \| Q_i\right) = \sum_i \sum_j p_{j|i} \log\left(p_{j|i} / q_{j|i}\right) \tag{4}$$

In which P_i represents the conditional probability distribution over all other data points given data point x_i, and Q_i represents the conditional probability distribution over all other map points given map point y_i. SNE constructs reasonably good visualizations, however, it is hampered by a cost function that is difficult to optimize and by a "crowding problem" as named in [6]. Maaten and Hinton adapted SNE by replacing conditional probability distribution $p_{j|i}$ by joint probability distribution p_{ij}, and using a student t-distribution with one degree of freedom as the heavy-tailed distribution in the low dimensional map. The joint probabilities q_{ij} is defined as

$$q_{ij} = \left(1 + \left\|y_i - y_j\right\|^2\right)^{-1} \Big/ \sum_{k \neq l} \left(1 + \left\|y_k - y_l\right\|^2\right)^{-1} \tag{5}$$

The gradient of the Kullback-Leibler divergence between two joint probability distributions, P_i in high dimensional space and Q_i in low dimensional space, is given by

$$\frac{\delta C}{\delta y_i} = 4 \sum_j \left(p_{ij} - q_{ij}\right)\left(y_i - y_j\right)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1} \tag{6}$$

Implementation of t-SNE with MATLAB code can be found in [8]. Performance comparison of t-SNE with a DBN configuration proposed by Hinton in [5] strengthens our belief on t-SNE. We tend to regard t-SNE as a convenient tool for DBN based classifier, i.e., we can check the separability of samples subjectively by watching their t-SNE visualization before designing a DBN-based classifier for them. If data are separable by t-SNE visualization, then it's OK to go for DBF training. Otherwise, it may be more reasonable to peruse the sample representations for better separability.

4 Experiments and Results

Voices from two storytellers were selected as clean speech. One storyteller is female (Liu Lanfangj), and another one is male (Yuan Kuocheng). Their voices provide sufficient clean speech for the study. Babble noise was selected as the disturbance signal. The noisy speech is combined by clean speech plus level adjusted noise. Noise level was adjusted according to the preset global SNR values, i.e., -5dB, 0dB and 5dB in the study. Each speaker contributed about 25 minutes speech which created 85000 samples as a dataset. Following descriptions only refer to the female's samples.

For the purpose of training DBN as a binary classifier for AMS-SNR pairs, number of samples in each class is expected to be equal. It requires a separation of the sample SNR values with a threshold. Two fixed thresholds (-8dB for subband 1-15 and -16dB for subband 16-25) were used in [3]. This is inappropriate for equal division of the samples.

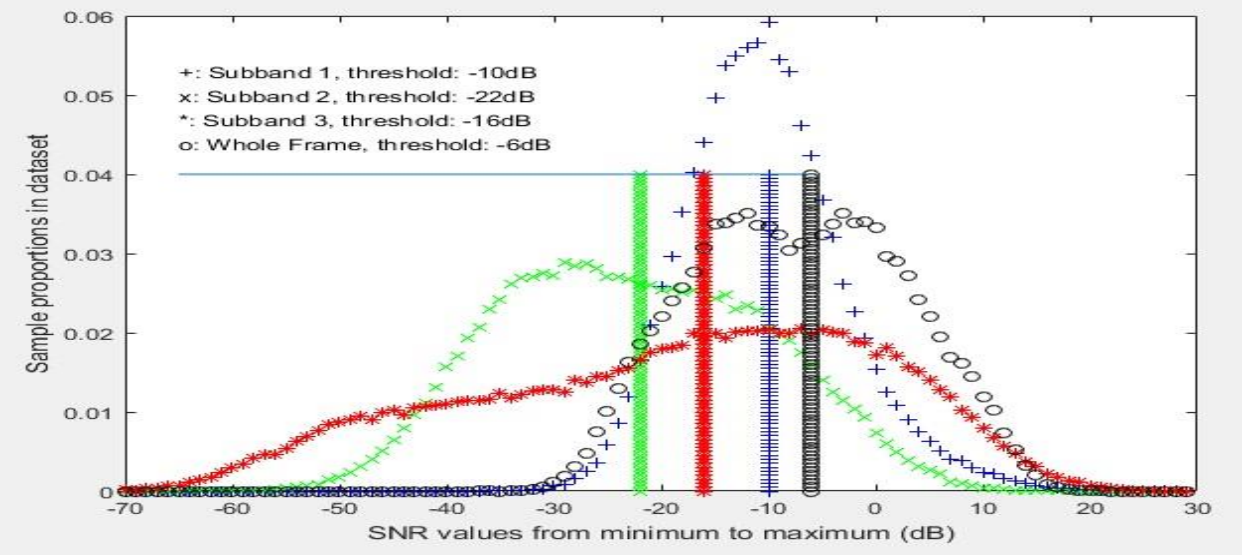

Figure 2. Distribution of sample proportions versus SNR values (dB)

Fig. 2 is the distribution of sample proportions versus SNR values and the threshold for half to half separation of the samples in first three subbands. As a comparison, SNR values corresponding to noisy speech before subband filtering are also shown in the figure (labeled as "Whole Frame"). It is obvious that thresholds in different subbands differ from each other drastically. Table 1 lists the detailed range of all SNR values in each subband and SNR thresholds for samples to be separated half to half.

Table 1. Range of local SNR values in each subband and the thresholds for samples to be separated half to half (dB)

subband No.	minSNR	maxSNR	threshold	subband No.	minSNR	maxSNR	threshold
1	-35	28	-10	14	-62	34	-16
2	-66	19	-22	15	-61	41	-16
3	-73	29	-16	16	-56	37	-15
4	-73	34	-12	17	-56	34	-15
5	-66	37	-12	18	-54	34	-15
6	-62	35	-15	19	-60	34	-17
7	-65	35	-18	20	-59	37	-18
8	-75	34	-19	21	-60	34	-16
9	-71	38	-15	22	-51	32	-11
10	-61	37	-13	23	-49	32	-11
11	-56	40	-11	24	-53	27	-16
12	-57	42	-13	25	-59	21	-21
13	-63	38	-14	Whole Frame	-36	22	-6

According to Table 1, the thresholds which separate samples equally in different subbands change from -22dB to -10dB and have a continuous distribution. It is our opinion that thresholds in different subbands had better be set individually. This will insure, at least at the overall level the same sample number of two classes.

Fig.3 displays the effect of AMS feature vectors visualized by t-SNE with 2000 samples in subband 1. AMS feature vectors were firstly labeled as class '0' or class '1' samples according to their SNR values. These samples were then processed by principal component analysis (PCA) method. This routine adjusted sample dimension to appropriate size. Next distance between any two samples was calculated, and the distance matrix was converted to joint probability distribution matrix. Affinity transformation of the resulted joint probability matrix implemented by t-SNE maps all the samples to a two dimensional plane. They were labeled by '0' or '1' as shown in Fig.3.

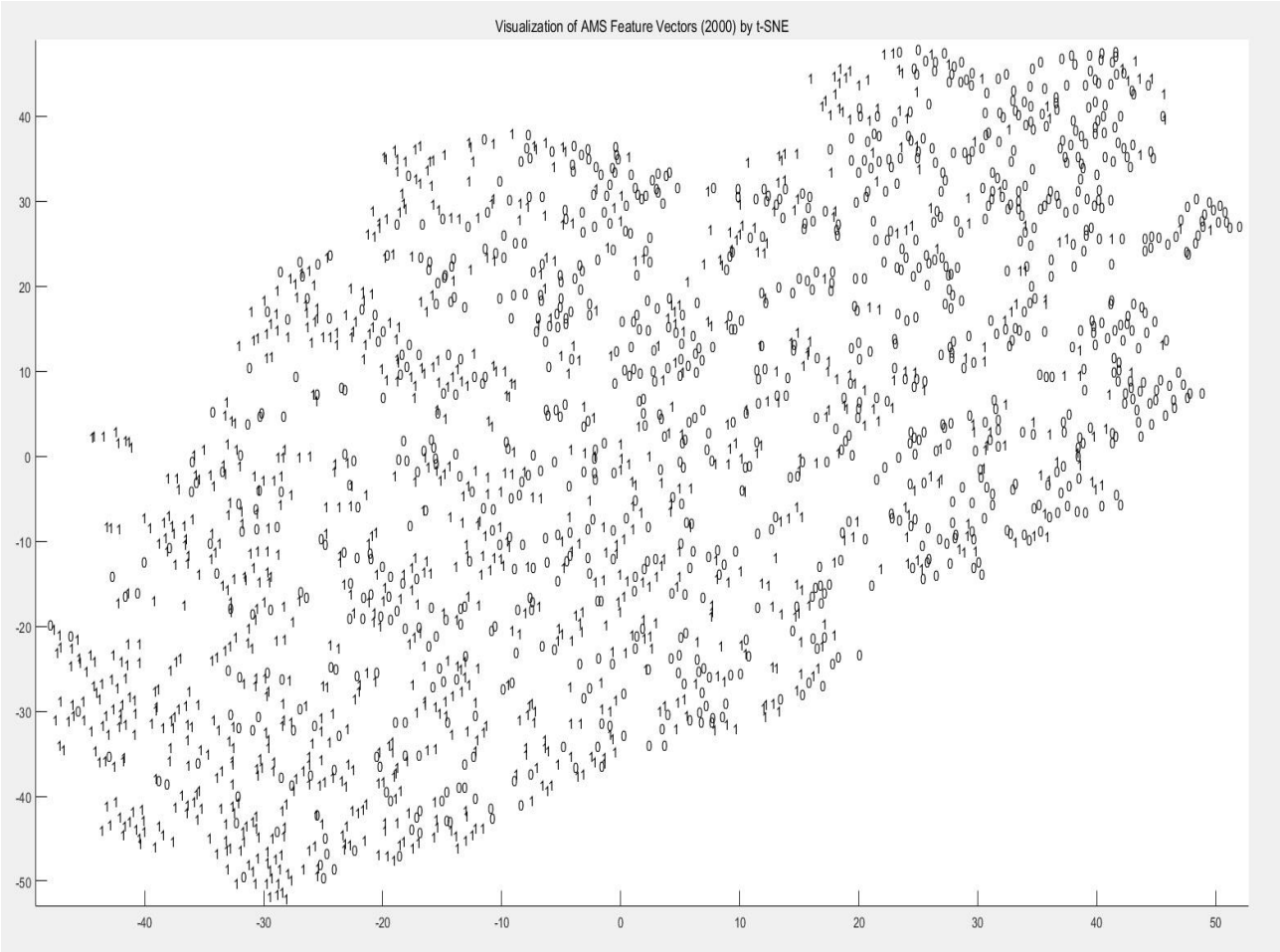

Figure 3. Visualization of AMS feature vectors by t-SNE, 2000 samples

The Separability in Fig.3 is not clear as much as expected. However, it does show some interesting information. First of all, class '0' and class '1' have their dominant zones respectively. The potential situation might be they are separable. Secondly, confused zones with '0' and '1' are very similar to each other. It might be associated with certain unknown or ignored reasons. Since the computational complexity increases at square rate in t-SNE, it is difficult to break the code with much more samples. Our preliminary judgement is that AMS is an effective representation of SNR values in speech. It is appropriate to be employed in DBN based classification.

5 Conclusions

The separability of local SNR values represented by AMS feature vectors was studied in this paper. The t-SNE technique was introduced to help demonstration of samples separability. Although t-SNE shows good performance in visualization of high dimensional data, it is not a classifier. Our research aims to train DBN with AMS feature vectors, and this study reinforces the determination of focusing on DBN configuration. Clues resulted from this work have led to research on new topics, e.g., AMS samples cleaning, sample size equalization between intervals with same length, etc. The latter refers to the fact that fine grain SNR value estimation needs relevant sample structure. Not only equal sample number in overall, but also approximate same sample number in each interval. Currently, research is focused on the accuracy of the binary classification with DBN.

Acknowledgment

This work was funded by the Natural Science Foundation of Shaanxi Province under the Grant No. 2014JM2-6121. It is also supported by Xi'an Peihua University under the 2016 authorized project list classified as major item in natural science study. The authors would like to thank funds from both departments, as well as helpful opinions from colleagues.

References

[1] W. Wei, Y. P. Chen, "Speech Enhancement by Spectral Component Selection", 5th International Conference on signal processing proceedings, Volume 2, pp.674-678, August 21-25, 2000, Beijing, China

[2] W. Wei, Q. S. Xie, Q. J. Chen, "SNR Classification Based on Amplitude Modulation Spectrogram via Deep Belief Networks", Proceedings of IMCEC 2016, pp.1834-1839, Oct. 3-5, 2016, Xi'an, China

[3] G. Kim, Y. Lu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," The Journal of the Acoustical Society of America, 126(3), pp1486-94, Sep. 2009

[4] J. Tchorz and B. Kollmeier, "SNR Estimation Based on Amplitude Modulation Analysis With Applications to Noise Suppression," IEEE Trans. on Speech and Audio Processing, Vol.11, No3, pp.184-192, May 2003

[5] G. Hinton, and R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," Science, Vol. 313, pp.503-507, Jul. 2006

[6] L. Maaten, G. Hinton, "Visualizing Data using t-SNE", Journal of Machine Learning 9 (2008), pp.2579-2605.

[7] G. Hinton and S. Roweis, " Stochastic Neighbor Embedding". In Advances in Neural Information Processing Systems, volume 15, pages 833–840, Cambridge, MA, USA, 2002. The MIT Press

[8] https://lvdmaaten.github.io/tsne/, Toolbox for t-SNE MATLAB implementation