Microphone Array Speech Denoising Modeled by Tensor Filtering

Jing Wang, Yahui Shan, Shequan Jiang, Xiang Xie

School of Information and Electronics, Beijing Institute of Technology

wangjing@bit.edu.cn

Abstract Analysis of the customers' satisfaction provides a strong guarantee to improve the service quality in call centers. In this paper, an intelligent satisfaction recognition system is introduced to analyze the customers' attitude after the dialogue through the emotion recognition of recording. It is a two-layer model to accomplish satisfaction analysis. In the first layer, we mainly detect the customers' emotion by their voice, and in the second layer we extract 54 satisfaction features based on emotion recognition results and establish the mapping model between emotion and satisfaction. SVM and ELM is used as the mapping model. According to the experiment, SVM has the best performance of F score, which is 0.71 and the training time is 3155 seconds. Compared to SVM, the maximum F score of ELM is boosted up to 0.723 and training time is reduced to 7.28 seconds.

Introduction

Speech enhancement technology has great importance in voice communication with the purpose of improving the speech quality under noisy environment and increasing the performance of subsequent processing system. By means of installing multiple microphones on the front-end in speech processing system, there are various means to overcome the adverse impact of the background noise, and it can improve the quality of speech communication and speech recognition. This paper proposes to model the multi-microphone speech signal into a 3-order tensor form according to three dimensions of channel, time and frequency and to carry out the speech denoising based on the tensor algebra analysis theory. We build multi-mode linear filters to reduce noise according to low rank tensor approximation with Tucker decomposition and alternating least square algorithm. On the whole, the proposed tensor filtering method can perform well and show the potential ability in our experimental simulation.

Tensor Decomposition

There are many choices for tensor decompositions which generally combine a choice of orthonormal bases in the domain of the tensor with a suitable truncation of its expansion. Two main kinds of tensor decompositions are CP (CANDECOMP/PARAFAC) and Tucker decomposition. The latter one can be regarded as a multilinear generalization of the traditional matrix SVD (Singular Value Decomposition) and plays an important role in tensor-based signal processing.

For an N-order tensor, a low rank approximation with the truncated Tucker decomposition can be represented as

$$\boldsymbol{X} \approx \boldsymbol{G} \times_1 \boldsymbol{U}^{(1)} \times_2 \boldsymbol{U}^{(2)} \circ \circ \circ \times_N \boldsymbol{U}^{(N)}$$

Where , are the truncated components or factor matrices (usually orthogonal matrices) in mode-1, mode-2 and mode-n subspaces, respectively. is the core tensor whose entries show the level of interaction between the different components, and is computed with

$$\boldsymbol{G} = \boldsymbol{X} \times_1 \boldsymbol{U}^{(1)\mathrm{T}} \times_2 \boldsymbol{U}^{(2)\mathrm{T}} \circ \circ \circ \times_N \boldsymbol{U}^{(N)\mathrm{T}}$$

The Tucker decomposition can be realized by higher-order SVD (HOSVD) method. One advantage of Tucker decomposition is that it can transform the original tensor into the core tensor with factor matrices. It is very useful in low rank approximation and dimensionality reduction.

Experiment Design

The microphone array database used in the experiment comes from Carnegie Mellon University (CMU). This database is recorded in a real environment and contains 10 male speakers each speaking 14 utterances. There are two kinds of microphone array including 8 elements and 15 elements. We used the 8-element array microphones which are spaced linearly with a spacing of 7 cm between elements. The speaker sit directly in front of the array at a distance of 1 meter from the center and the noise mainly contains disk-drive and cooling fans of many computers in the real environment. The data sampling rate is 16 kHz with quantization precision of 16 bits PCM. We choose 15 utterances randomly from the database. Hamming window is chosen as frame window of length 32ms with 50% overlapping. Taking the utterance 'beeoer' as an example, the original dimensions of the noisy tensor signal $\boldsymbol{T} \in \boldsymbol{R}^{Ic \times It \times Is}$ are set as I_c=8, I_t=150 and I_s=512.

In this paper, the noise **N** is assumed to be independent from the clean signal **S** and the n-mode rank K_n of tensor **T** is less than the n-mode dimension I_n. The target clean speech tensor will be estimated from the different signal subspace of the noisy tensor data using low rank approximation method. In the low rank tensor approximation, the key problem is to find the optimal parameters K_n which finally affect the performance of the speech denoising. Here we estimate the optimal K_n by the use of MDL (Minimum Description Length) criterion. The optimal signal subspace dimension is just the optimal rank under each mode.

From Figure 1, we can see that the minimum value of the MDL estimation is 7, 94 and 74 respectively. Thus for the different subspace, the optimal tensor rank is set to be K_1=7, K_2=94 and K_3=74. MDL criterion can effectively compute the optimal dimension of the subspace instead of selecting fixed dimension artificially. It can make the best compromise between signal distortion and noise reduction.

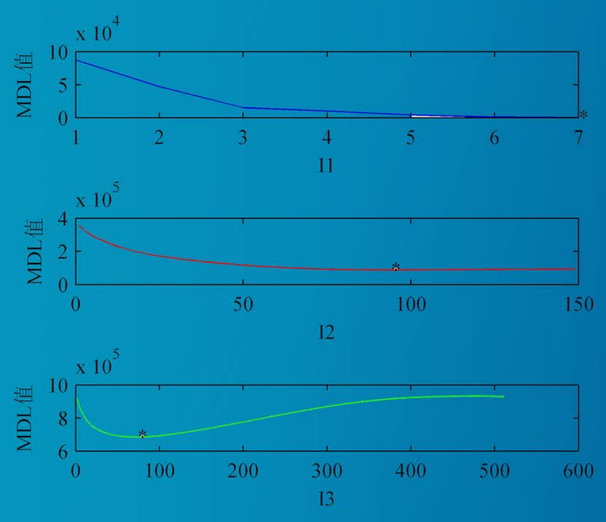

Figure 1: Estimation results of tensor rank with MDL

Result

Several objective speech quality evaluation indexes, include segmental signal-to-noise ratio enhancement (SNRseg), log-likelihood ratio (LLR), overall quality (Covl), background distortion (Cbak) are used to compare the performance of different speech denoising algorithms.

In Table1, "Noisy input" represents the middle microphone speech signal of the microphone array. All scores are obtained with an average of the 15 chosen utterances. Note that higher SNRseg, Covl, Cbak scores and smaller LLR score represent better objective performance.

Table 1. Objective experiment results.

Method	SNRseg	Cbak	Covl	LLR
Noisy input	--	2.014	2.959	0.724
GSC	1.002	2.127	2.967	0.752
Subspace	2.566	2.165	3.057	0.774
Tensor Filter	1.305	2.230	3.002	0.840

The objective results indicate that the noise reduction degree and the spectrum similarity are both lower for the tensor filtering method which can be further improved by considering the subspace signal feature.

From Table 2, we can see that most of the listeners considered that the subspace approach can remove most of the noise signal and lead to the least residual which has a consist result with the objective 'SNRseg' index.

Table 2. Subjective A/B test results.

Method	Least residual noise (%)	Least signal distortion (%)
GSC	9.17	15.00
Subspace	45.00	15.83
Tensor Filter	16.67	22.50

While the tensor filtering method has a better result with the subjective feeling of signal distortion mostly because of the accurate rank estimation in each subspace.

Conclusions

This paper constructs the noisy multi-microphone speech data using 3-order tensor model with three dimensions: channel, time, frequency. The traditional wiener filtering can be processed by tensor filtering model which is realized by low rank tensor approximation with Tucker model. This kind of higher-order speech denoising method based on tensor model extends the speech signal processing approaches in high dimension. Preliminary results show the potential improvement on the denoised signal quality by use of tensor filtering method. While the method' s noise reduction ability performs worse and should be enhanced if compared with the subspace approach. The research of this paper can provide some useful results of tensor filtering in microphone array speech denoising. In order to get a better noise reduction, the tensor constructing and filtering method could be improved referring to the processing scheme of the subspace approach.