Discrimination of Speech and Ship-radiated Noise Based on Frequency Spectrum Similarity

LI Dawei, YANG Rijie, Han Jianhui
Department of Electronic and Information Engineering
Naval Aeronautical and Astronautical University
Yantai, China
e-mail:latt68@sina.com

***Abstract*—In this paper, the discrimination of speech and ship-radiated noise has been studied, for their interference with each other in the modern sea war, and a new feature abbreviated as SFSD (the Similarity of the Frequency Spectrum Distribution of two successive signal frames) is introduced. Based on the feature, a new two-stage approach is proposed, which firstly computes the SFSF and classifies the signal into speech or ship-radiated noise roughly, and then in the second stage, with the help of context smoothing, the signal are final discriminated in to speech or ship-radiated noise. The algorithm is benchmarked on a large measured data set, with correct recognition accuracy of 96% for ship-radiated noise and that of 92% for speech. Experimental results show that efficiency is exceptionally good.**

Keywords—discrimination; similarity; frequency spectrum; two-stage approach;context smoothing;

I. INTRODUCTION

With the use of all kinds of communication equipments in the modern sea war, mutual Interferences have been hard to avoid, and one of which is the interference with each other of speech signals and ship-radiated noise. So, it is necessary to discriminate speech signals with ship-radiated noise correctly, but to now few studies have been done on it.

Fortunately the problem is close to the audio discrimination tasks, such as the speech/music discrimination, and some feature extraction and classification techniques are instructive to our task.

Zhang [1] used features like the energy function, average ZC rate, the fundamental frequency and the spectral peaks tracks, and classified the signals into music, speech, song, environmental sound and silence, etc. They achieved an accuracy of more than 90% in audio segmentation.

Cepstral coefficients and Gaussian mixtures are used by Moreno and Rifkin [2] to model data and a support vector machine is trained for the classification, a performance of 81.8% is reported.

Tzanetakis [3] proposed a real-time classification method using three feature sets (timbre texture, rhythmic content and pitch content). their music/speech classifier has 86% accuracy.

Cooper [4] focus on the similarity matrix, a two-dimensional matrix which measures the (dis)similarity between any two instances of the audio, and it can be used in such tasks as audio segmentation, music summarization and beat estimation, and our similarity analysis.

And in many other methods, spectral entropy, sample entropy, time-frequency analysis [5], histogram equalization-based features [6] and local discriminating bases (LDB) technique [7] are also used.

The above-mentioned literature show that the general method involves extracting a large number of features and feeding them to a pattern classifier. However, the features extraction is one of the most computationaly expensive tasks and including too many features could provide no benefit if some features are correlated with others.

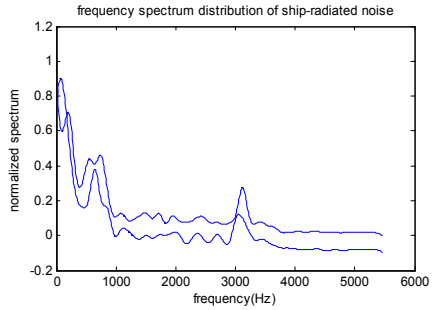
So, in this paper, only the frequency spectrum is focused on and a new feature abbreviated as SFSD (the Similarity of the Frequency Spectrum Distribution of two successive signal frames) is introduced. And based on the feature, a two-stage discriminating method is proposed, in which the two signals are discriminated roughly in the first stage using SFSD, and then in the second stage, with the help of context smoothing, the final discrimination is realized. Experimental results show that the efficiency is exceptionally good.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

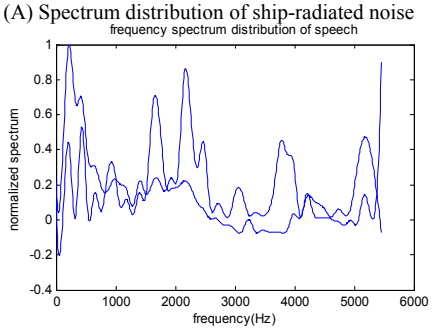
II. DESCRIPTION OF SIGNAL CHARACTERISTICS

Ship-radiated noises are mainly generated from the work of mechanical equipments of ships, such as the propeller rotation and so on. That is to say, for a ship, its shape, propeller, structure and so on are of the key factors that affect the frequency spectrum of its radiated noise. And in the real condition, within a very short interval, all parts mentioned above change very little and that results in the key factors keep stable. So, the frequency spectrum of the noise will keeps stable within a very short interval such as one second whatever the ship is in any kinds of sailing conditions, thus the frequency

This work was supported in part by the National Natural Science Foundation of China (61271444)



(A) Spectrum distribution of ship-radiated noise



(B) Spectrum distribution of speech signal

Fig. 1. Frequency spectrum of two successive signal frames.

spectrums from successive frames will change very little too. Fig.1 (A) shows what has been discussed above. The two curves are calculated form two successive frames in 20-ms duration, and each is normalized. As we analyzed above, the two curves change very little.

But it is quite different for speech signals, just like what shown in Fig.1 (B). For speech signals, phonemes are the minimum pronounce unit and each of them has its unique and distinct articulation, forming its distinct frequency spectrum. Phonemes constitute words expressing the meaning of the speech signals, and thus, the word that consists of several phonemes will show its distinct frequency spectrum which is different from that of other words with different meaning. So even in a very short interval which is equal to the interval mentioned in ship-radiated noises analysis, the two spectrum curves from successive frames change a lot for the reason of the different phonemes in each frame. Just as what is shown in Fig.1 (B), the change of the two curves which is calculated in the same way as the curves in Fig.1 (A) is very obvious.

So, from Fig.1, it can be found that frequency spectrum of two successive frames of speech signal and ship-radiated noise is obviously different, and the difference can be exploited to discriminate speech and ship-radiated noise. Here, the variance of the SFSD (V_{sfsd}) is calculated.

III. Discrimination Algorithm

In this section, the discrimination method is implemented based only on SFSD. For each signal segment, the values of SFSD can be acquired after applying FFT to each frame. And then their variance V_{sfsd} can be calculated.

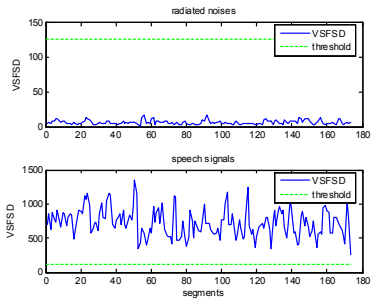


Fig. 2. V_{sfsd} distribution and it is calculated in 1-s segment

A. Feature Extraction

The feature is based on a similarity measure, and a known similarity measure defined on the probability density function is used here:

$$\rho(p1,p2)=\int\sqrt{p1(x)\cdot p2(x)}dx \tag{1}$$

Given a matured signal segment x, the Short-Time Fourier Transform $F_i(w)$ $(i=0,1,\mathrm{L}\ 49)$ is calculated for each frame in this segment, using Hamming windows of n samples, and then the magnitude spectrums $f_i(w)$ $(i=0,1\mathrm{L}\ 49)$ are derived by taking the absolute value of the elements of $F_i(w)$, after discarding the symmetric part. And then, for the discrete magnitude spectrum $f_i(w)$ and $f_{i+1}(w)$ of every two successive frames in this segment, the similarity measure based on (1) can be expressed as follows:

$$SFSD_i[f_i,f_{i+1}]=\frac{\sum_{w=0}^{n/2}[f_i(w)\cdot f_{i+1}(w)]}{\sqrt{\sum_{w=0}^{n/2}[f_i(w)]^2}\cdot\sqrt{\sum_{w=0}^{n/2}[f_{i+1}(w)]^2}} \tag{2}$$

Then, the variance of these $SFSD_i$ is calculated because of that V_{sfsd} is more stable than using these $SFSD_i$ features directly. As shown in Fig.2, V_{sfsd} is sufficiently different between speech segments and radiated noise segments. And there is an obvious boundary.

B. Threshold Determination and Context Smoothing

To feed the need of real-time processing, some pattern classifier should be avoided and the feature is limited to be used only the V_{sfsd}. So the threshold and the classifying method are obviously significant.

1) Multiple thresholds selection

The threshold determination is based on the histogram of the V_{sfsd} distribution.

$$His(d)=\frac{n_d}{n}\quad d=0,1\mathrm{L}\ d_{\max}-1 \tag{3}$$

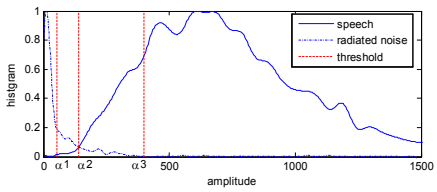


Fig. 3. The histogram of V_{sfsd} from speech and radiated noise.

Where n_d is the number of V_{sfsd} whose value is equal to d, and n is the total number of V_{sfsd}. Fig.3 shows the histogram distribution for speech signal and radiated noise. After experiments based on a large measured data set, the three values $\alpha 1 = 50$, $\alpha 2 = 145$ and $\alpha 3 = 405$ is selected as the multiple thresholds and context smoothing is introduced to improve the discrimination accuracy.

2) context smoothing

speech and radiated noises normally last periods of time, so, if there is a sudden change in some segment t but its neighbor segments $t+1$ and $t-1$ are the same categories, the change is in most cases caused by some instant noise and should be ignored except that V_{sfsd} of current segment is large enough. The context smoothing process is as follows,:

When $V_{sfsd} \geq a1$, the segment is discriminated to be speech.

When $V_{sfsd} \leq a1$, the segment is discriminated to be noise.

When $\alpha 1 < V_{sfsd} \leq \alpha 2$, the segment is discriminated to be undetermined radiated noise signals.

When $\alpha 2 < V_{sfsd} < \alpha 3$, the segment is discriminated to be undetermined speech signals.

Context smoothing process is only used to the last two undetermined segments. This procedure introduces a delay of 1 s, which is necessary for the final determination. At the end of this procedure, the category of each segment is decided.

C. Discrimination Procedure of the Algorithms

The discriminator of speech and ship-radiated noises consists of a sequence of tests based on the value of V_{sfsd} and the procedure can be described as follows:

Step 1: Calculate the frequency spectrum of each frame in the given segment by applying a STFT with hamming window. And to improve the performance of the algorithm, each of the distributions can be fitted by some fitting method.

Step 2: Calculate the values of V_{sfsd} based on the similarity measure.

Step 3: Determine the category of the segment according to the given three thresholds and with the help of the context smoothing method

After the three steps above, the speech signals and the ship-radiated noises can be discriminated correctly, and only a little calculation amount is needed, which meets the need of real-time processing without sacrificing performance.

IV. Experiments

The proposed algorithm are tested on a large measured data set containing speech signals spoken by a variety of both male and female speakers and ship-radiated noise including many kinds of ships sailing in various sea condition. Both of the data are 400 minutes and their sample frequency is 11025Hz.

The experiment is divided into three parts to assess the performance of the algorithm, as follows:

Data set 1: This data set consists of speech records from male and female speakers at different ages, and 20 seconds segments of each record are intercepted. This set is mainly used to test if the detection algorithm is applicable to all kinds of speech signal and the experimental result is shown in Fig.4.

Data set 2: This data set consists of different ship-radiated noise records, and its structure is the same as data set 1 except that this set is used to test the feasibility of the algorithm to different kinds of strong background of ship-radiated noise. Its result is shown in Fig.5.

Data set 3: This data set is very important for the algorithm test because of that it consists of records coming from different speakers in different background of strong ship-radiated noise. One result is shown in Fig.6.

Each figure contains four plots: the first subplot is the final discrimination, where 1 corresponds to radiated noise and 4 corresponds to speech signals; the second subplot is the result without context smoothing, where 1 and 4 is the same as that in the first subplot and 3 corresponds to undetermined speech and 4 corresponds to undetermined speech signals, the third subplot is the V_{sfsd} distribution, and the fourth subplot is the signal amplitude. We can see from these figures that the discriminator is available to all kinds of ship-radiated noises and speech signals.

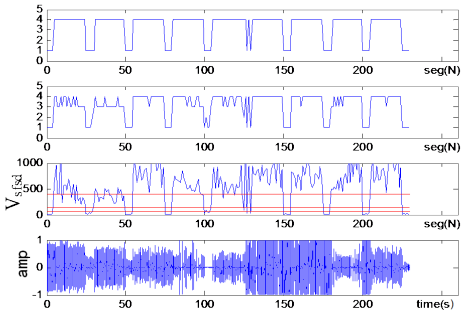


Fig. 4. Discrimination results based on the data set 1

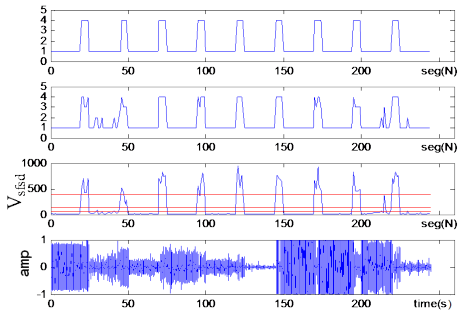


Fig. 5. Discrimination results based on the data set 2

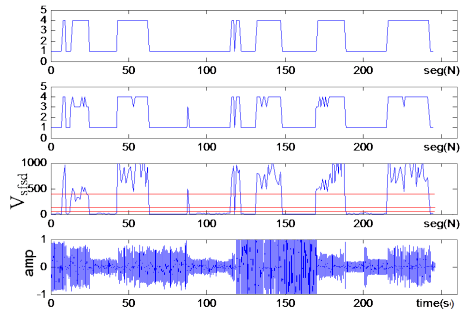


Fig. 6. Discrimination results based on the data set 3

$$\Delta = \frac{DurationofCorrectlyDiscrimiated}{DurationofTotalSegments} \tag{4}$$

The effectiveness of the algorithm is further tested in terms of the percentage of the signal detected correctly using (4), and the detection accuracy of the proposed algorithm is shown in table 1.

Table 1 shows the discrimination accuracy of the three experiments using (4), and we can see that the algorithm achieves high consistent accuracies. The lower accuracy of the speech signals are mainly because of the more pause or quiet intervals between words which usually have a low value of V_{sfsd} too. But the incorrect discriminations of the pause and quiet segments in speech signals have little effect to the whole task of discrimination processing.

All the experimental results show that the proposed algorithm using SFSD is exceptionally good for our task and the detection accuracy is acceptable.

TABLE I. DISCRIMINATION RESULTS OF THE THREE EXPERIMENTS

Data Sets	True Speech(%)	True Noise(%)	Accuracy(%)
1	93.25	97.32	94.06
2	91.74	96.60	95.63
3	92.59	98.43	95.51

V. *CONCLUSION*

In this paper, an effective algorithm for speech detection in background of strong ship-radiated noise is presented using a simple feature REST calculated based on spectrum entropy. The algorithm is tested on many kinds of measured data sets and the experiments results illustrate the efficiency of the algorithm. We can conclude that the REST is capable of working well for speech detection in the strong background of ship-radiated noise.

References

[1] T. Zhang and J. Kuo, "Audio content analysis for on-line audiovisual data segmentation and classification," IEEE Trans. Speech Audio Process., vol. 9, no. 3, pp. 441–457, May 2001.

[2] P. Moreno and R. Rifkin, "Using the fisher kernel mothod for web audio classification," in Proc. IEEE Conf. on Acousitcs, Speech, and signal Processing(ICASSP), 2000, pp.1921-1924.

[3] Tzanetakis, Cook, "Musical genre classification of audio signals," IEEE Trans speech audio process, Vol.10, no.4, pp293-302, Jul.2002.

[4] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in Proc. 3rd Int. Conf. Music Inf. Retrieval, Paris, France, Oct. 13–17, 2002, pp. 81–85.

[5] Karthikeyan Umapathy, Sridhar Krishnan and Shihab Jimaa, "Multigroup classification of audio signals using time-frequency parameters," IEEE Transactions on multimedia, vol. 7, no. 2, pp.308-315, April. 2005.

[6] Ascension Gallardo-Antolin and Juan M. Montero, "Histogram Equalization-based features for speech, music, and song discrimination," IEEE signal Processing letters, vol. 17, no. 7, Jul. 2010.

[7] Karthikeyan Umapathy, Sridhar Krishman and Raveendra K. Rao, "Audio signal feaure extraction and classification using local discriminant Bases," IEEE Transation on audio,speech and language processing, vol.15, no.4, pp.1236-1246, May.2007

[8] S. Vembu and S. Baumann, "Separation of vacals from polyphonic audio recording," in proc. 6th Int. Conf. Music Inf. Retrieval, London U.K. Set.11-15,2005, pp.337-344.

[9] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separation a foreground singer from background music," in Proc. Int. Symp. Frontiers of Res. Speech and Music, Mysore, India, May8-9, 2007.

[10] A. Ozerov, P. Philippe, f. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," IEEE Trans. Audio, Speech, Lang. Process., vol.15,no.56, pp.1564-1578, Jul.2007

[11] A. Pikrakis, I. Antonopoulos, and S. Theodoridis, "Music meter and tempo tracking from raw polyphonic audio," in Proc. 9th Int. Conf. Music Inf. Retrieval, Barcelona, Spain, Oct. 10–14, 2008.

[12] K. Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," EURASIP J. Adv. Signal Process. vol. 2007, no. 1, pp.1–11, Jan. 2010.

[13] M. A. Bartsch, "To catch a chorus using chroma-based representations for audio thumbnailing," in Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust., New Paltz, NY, Oct. 21–24, 2001, pp. 15–18.

[14] Costas panagiotakis and George Tziritas, "A speech/music discriminator based on RMS and zero-crossings," IEEE transactions on multimedia, vol.7, no.1, pp.155-167, Feb.2005.

[15] saac Álvarez, Luz García, Guillermo Cortés, Carmen Benítez, and Ángel De la Torre, "Discriminative Feature Selection for Automatic Classification of Volcano-Seismic Signals," IEEE Geoscience and remote sensing letters. vol.9,No.2, Mar 2012.