Journal of Southeast University (English Edition) Vol. 28, No. 3, pp. 261 - 265 Sept. 2012 ISSN 1003—7985

Whisper intelligibility enhancement based on noise robust feature and SVM

Zhou Jian[1,2] Zhao Li[1] Liang Ruiyu[1] Fang Xianyong[2]

([1]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2]Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Anhui University, Hefei 230601, China)

Abstract: A machine learning based speech enhancement method is proposed to improve the intelligibility of whispered speech. A binary mask estimated by a two-class support vector machine (SVM) classifier is used to synthesize the enhanced whisper. A novel noise robust feature called Gammatone feature cosine coefficients (GFCCs) extracted by an auditory periphery model is derived and used for the binary mask estimation. The intelligibility performance of the proposed method is evaluated and compared with the traditional speech enhancement methods. Objective and subjective evaluation results indicate that the proposed method can effectively improve the intelligibility of whispered speech which is contaminated by noise. Compared with the power subtract algorithm and the log-MMSE algorithm, both of which do not improve the intelligibility in lower signal-to-noise ratio (SNR) environments, the proposed method has good performance in improving the intelligibility of noisy whisper. Additionally, the intelligibility of the enhanced whispered speech using the proposed method also outperforms that of the corresponding unprocessed noisy whispered speech.

Key words: whispered speech; intelligibility enhancement; noise robust feature; machine learning
doi: 10.3969/j.issn.1003-7985.2012.03.001

Recently, processing of whispered speech has received much attention. As a special paralinguistic phenomenon in human communications, the whisper has two major characteristics in contrast to phonated speech. The first difference is the lack of pitch and turbulence like noise excitation patterns since whispered speech is produced with no vibration of the vocal cords. The second difference is the coupling of the trachea with the vocal tract due to the opening of the vocal folds[1].

Received 2012-04-26.
Biographies: Zhou Jian (1981—), male, graduate, lecturer; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.
Foundation items: The National Natural Science Foundation of China (No. 61231002, 61273266, 51075068, 60872073, 60975017, 61003131), the Ph.D. Programs Foundation of the Ministry of Education of China (No. 20110092130004), the Science Foundation for Young Talents in the Educational Committee of Anhui Province (No. 2010SQRL018), the 211 Project of Anhui University (No. 2009QN027B).
Citation: Zhou Jian, Zhao Li, Liang Ruiyu, et al. Whisper intelligibility enhancement based on noise robust feature and SVM[J]. Journal of Southeast University (English Edition), 2012, 28(3): 261 - 265. [doi: 10.3969/j.issn.1003-7985.2012.03.001]

In comparison with voiced speech, whispered speech enhancement is clearly more difficult due to its acoustically turbulent noise and has much weaker energy than the phonated speech. As a result, whispered speech is more susceptible to interference[2]. Generally speaking, improving the intelligibility of the noisy whisper rather than the quality (e.g., SNR improvements or comfort of the enhanced speech) is much more important since semantic information retrieval becomes the dominating purpose in whisper communication.

However, the existing enhancement algorithms, such as the power subtraction method[3] and the log-MMSE algorithm[4], successfully improve speech quality, but fail to improve the intelligibility of noisy corpus because the objective function used in these algorithms aims to improve speech quality rather than speech intelligibility. The two distortions, e.g., speech distortion and noise distorting, produced in these algorithms are treated equivalently. However, these two distortions have different effects on improving speech intelligibility[5].

A human can successfully get the semantic information from the whisper in diverse noisy contexts. This remarkable ability is attributed to the human auditory perceptual system. A great achievement of the perceptual system is its auditory scene analysis (ASA) capability[6]. The ASA consists of two basic perceptual stages: segmentation and grouping. The segmentation stage decomposes an auditory scene into a collection of sensory elements in the joint time-frequency domain. Each of the sensory elements should primarily originate from a single sound source. The grouping stage aggregates the sensory elements that arise from the same source. Segmentation and grouping are governed by perceptual principles, or ASA cues, which reflect intrinsic sound properties, including harmonic, onset and offset, location, and prior knowledge of specific sounds[7].

In computational auditory scene analysis (CASA), ideal binary time-frequency masking (IBM) is a signal separation technique that retains mixture energy in time-frequency units where the local signal-to-noise ratio exceeds a certain threshold and rejects mixture energy in other time-frequency units. Wang et al.[8] and Li et al.[9] have also shown that multiplying the ideal binary mask with the noise-masked signal can yield large improvements in intelli-

gibility, even at extremely low SNRs of −10 to −5 dB.

In this paper, we explore this idea to improve the intelligibility of whispered speech in a noisy background. A novel feature called Gammatone feature cosine coefficients (GFCCs) based on an auditory periphery model is proposed and used to train an SVM classifier. The dominant whisper time frequency (T-F) unit is then recognized by a trained SVM classifier. The enhanced whisper is synthesized by multiplying the noisy whispered speech signal with the estimated binary mask.

1 Proposed Enhancement System

Fig. 1 shows the schematic diagram of the proposed system for enhancement of whispered speech. As shown in Fig. 1, the noise corrupted whisper is first preprocessed by cochlear filtering using a bank of 128 Gammatone filters to simulate the cochlea. In the feature extraction stage, the noise robust features called GFCCs are extracted for each time frame. In the training stage, noisy whispers are obtained by adding noise to the whisper manually with pre-prescribed SNRs. The local SNR of each T-F unit is first calculated and compared with the threshold (set to 0 in this paper) to obtain the prior binary mask, namely, the ideal binary mask (IBM). Then the extracted feature vector and the binary mask of each T-F unit are used to train an SVM classifier. In the enhancement stage, the features of the noisy whisper are extracted and given as input to the classifier. A binary mask estimated by the trained SVM is used to synthesize the whisper. In order to evaluate the performance of the proposed approach, listening tests are conducted to evaluate the intelligibility performance of whisper synthesized using the proposed system.

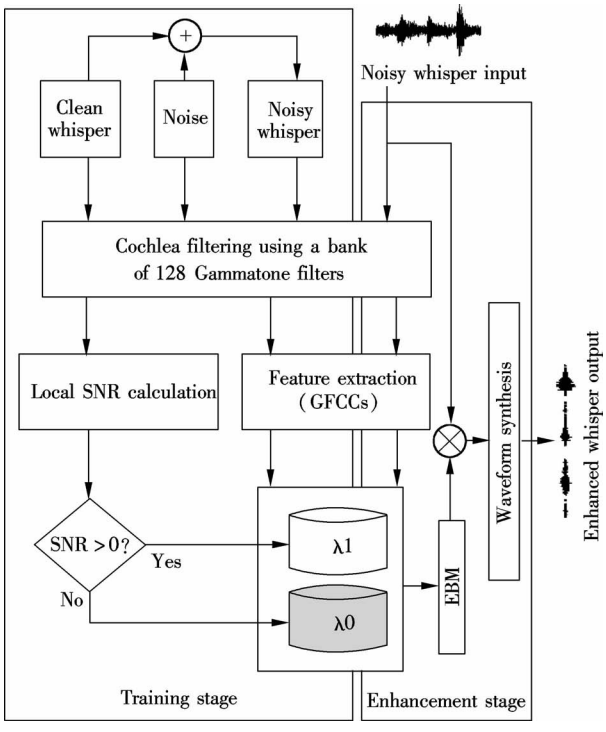

Fig. 1 Schematic diagram of the system for enhancement of whispered speech

2 Noise Robust Auditory Feature Extraction

A joint T-F representation of an input signal (with a 16 kHz sampling rate) is derived using a bank of Gammatone filters, which are derived from psychophysical observations of the auditory periphery. This filterbank is a standard model to simulate cochlear filtering. In this paper, the noisy whisper signal is first bandpass filtered into a bank of 128 Gammatone filters with center frequencies ranging from 80 to 8 000 Hz according to a mel-frequency spacing. This frequency range is adequate for speech understanding. The impulse response of a Gammatone filter centered at frequency f is

$$g(f, t) = \begin{cases} b^a t^{a-1} \mathrm{e}^{-2\pi bt} \cos(2\pi ft) & t \geqslant 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $a = 4$ is the order of the filter and b is the equivalent rectangular bandwidth, which increases with the center frequency f. Let $x(t)$ be the input signal. The response from a filter channel c is given by

$$x(c, t) = x(t) * g(f_c, t) \quad (2)$$

where "$*$" denotes convolution operation and f_c is the center frequency of filter channel c. The output signal from each channel is down sampled using low-pass filtering and decimation by a factor of 4, and then segmented into overlapping segments of 128 samples (32 ms) with an overlap of 64 samples, respectively. Each segment is Hanning-windowed and denoted by $\boldsymbol{A}_{t,c}$ which is referred to as a Gammatone feature (GF). Note that the dimension of a GF vector is much greater than that of feature vectors used in a typical speech recognition system. The Gammatone features are largely correlated with each other. In order to reduce the dimensionality and de-correlate the components, a discrete cosine transform (DCT) is applied to a GF as follows:

$$C_{t,c}(i) = \sqrt{\frac{2}{N}} \sum_{p=0}^{N-1} A_{t,c}(p) \cos\left(i\pi \frac{2p+1}{2N} \right) \quad (3)$$

where $i = 0, 1, \ldots, N-1$ and N is the segment length. The resulting coefficients are called GFCCs of frame t at channel c. When performing inverse DCT of GFCCs, we find that almost all the GF information is captured by including up to 30 coefficients. Hence, we use the 30-dimensional GFCCs as a feature vector for each time frame in this paper. The static GFCCs feature of frame t at channel c is

$$\boldsymbol{V}_{t,c} = \{C_{t,c}(0), C_{t,c}(1), \ldots, C_{t,c}(N-1)\} \quad (4)$$

In order to incorporate temporal information, delta co-

efficients are also calculated as

$$\boldsymbol{D}_{V_{t,c}} = \frac{\sum_{b=1}^{B} b(\boldsymbol{V}_{t+b,c} - \boldsymbol{V}_{t-b,c})}{2\sum_{b=1}^{B} b^2} \tag{5}$$

where b is a neighboring window index and B denotes the half window length and it is typically set to 2.

Fig. 2 shows a schematic diagram of enhanced whisper synthesized from different channels along with the estimated binary mask. In the enhancement stage, the contaminated whispered speech is first filtered into 128 bands. To remove across-channel differences, the output of each filter is time-reversed, passed through the filter, and reversed again. The filtered waveforms are windowed with a raised cosine window every 32 ms with an overlap of 16 ms between segments, and then weighted by the estimated binary mask. The binary masks are estimated by the two-class SVM classifier that has been trained in the training stage. Each T-F unit of noisy whispered speech is subsequently retained or eliminated by the estimated binary mask. The estimated target signal is then reconstructed by summing the weighted responses of the 128 filters.

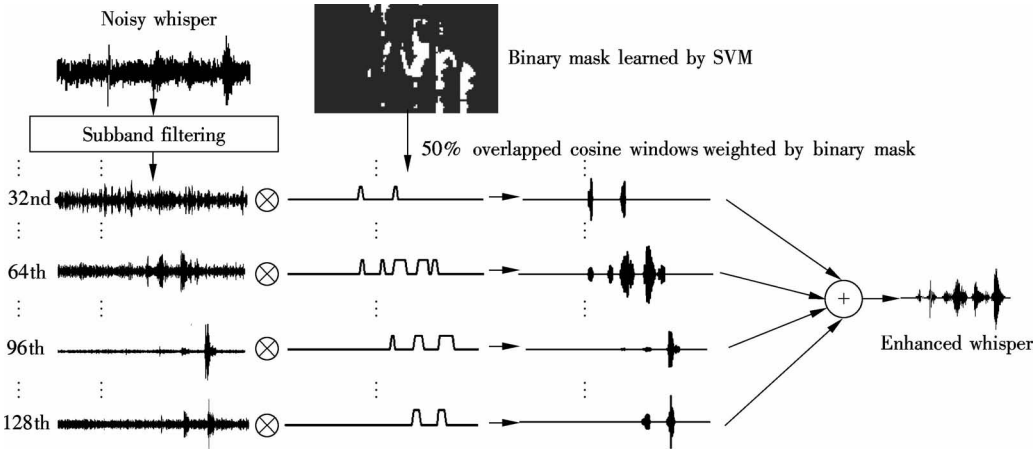

Fig. 2 Block diagram of the whisper enhancement stage of the proposed system

3 Experimental Results

In order to evaluate the proposed algorithm, whispered speech stimuli were collected. 50 phonetically balanced sentences were used to produce a whispered corpora. Each sentence was uttered by 2 male and 2 female speakers in a quiet room respectively. The format of the recordings is with a sampling rate of 16 kHz, 2 bytes/sample, and linear PCM. Each sentence was then artificially corrupted by noise at SNRs of −3, 0, 3 and 6 dB, respectively. Three types of noise recordings including Gaussian white noise, babble noise and car noise taken from NOISEX-92 database were used as noise maskers[10]. A noise segment of the same length as the speech signal was randomly cut out of the noise recordings and appropriately scaled to reach the desired SNR level. Therefore, a total of 2 400 (= 4 speakers × 50 sentences × 3 types of noise × 4 levels of SNR) noisy whispers were produced.

In the training stage, all the noise conditions (noise types: Gaussian, babble and car; SNRs: −3, 0, 3 and 6 dB) were used for training the SVM. Each T-F unit of a whispered speech from training set is first labeled as whisper dominated (denoted as 1) or noise dominated (denoted as 0). The feature vector of each T-F unit combined with the class label is used as the input of the classifier. The RBF (radial basis function) kernel is selected as the kernel function of the SVM. Unlike the linear kernel, the RBF kernel can handle the case when the relationship between class labels and attributes is nonlinear. The holdout cross validation method was repeated 10 times. The 9 of the 10 subsets were put together to form a training set and the remaining subset was used as the test set to estimate the binary mask each time.

Both the objective and the subjective evaluation performances of the proposed algorithm on speech intelligibility were conducted. The power subtraction and the log-MMSE algorithms were also evaluated and compared with the present algorithm. Five processing conditions, i. e., the noise-corrupted unprocessed whisper (denoted as UN), whisper enhanced using the ideal binary mask (denoted as IBM), whisper enhanced by power subtraction (denoted as PS), whisper enhanced by log-MMSE (denoted as log-MMSE) and whisper enhanced using the proposed algorithm (denoted as EBM), were considered.

A short-time objective intelligibility measure (STOI) proposed recently in Ref. [11] was used to compare the performance of the aforementioned algorithms in the aspect of intelligibility improvement. Fig. 3 illustrates the STOI values of the enhanced whispers by different algorithms in the context of babble noise and car noise with SNRs of −3, 0, 3 and 6 dB, respectively. It can be seen

from Fig. 3 that the proposed algorithm outperforms the other two algorithms even at very low SNR conditions. Results in Fig. 3 have also verified the conclusion proposed in Ref. [12] that the conventional algorithms do not necessarily improve the intelligibility of the enhanced speech in the low-SNR conditions.

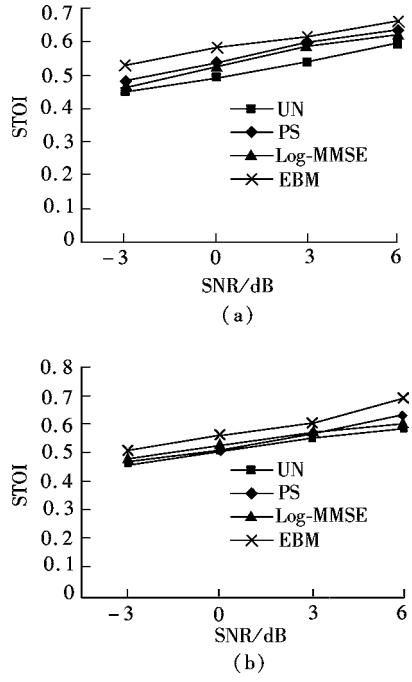

Fig. 3 Average STOIs of the unprocessed noisy whispers. (a) Babble noise environment; (b) Car noise environment

In the identification listening test, stimuli of UN, IBM and EBM were played to the listeners monaurally at a comfortable listening level. Listeners were asked to write down the words they heard. Intelligibility performance was assessed by counting the number of words identified correctly.

Fig. 4 shows the word identification rates of UN, IBM and EBM as the function of SNRs of −3, 0, 3 and 6 dB in different noise environments. We can see from Fig. 4 that the proposed system gains substantially higher intelligibility than the unprocessed stimuli in different noise environments with different SNRs. Its performance almost achieves the upper bound denoted by IBM.

Tab. 1 charts the mean Chinese word recognition rates of the whispers synthesized by the proposed method with different noises participating in the training stage. It can be seen from Tab. 1 that both the average word identification rates of the whispers enhanced by the power subtraction algorithm and the log-MMSE algorithm are lower than those of the proposed algorithm in different environments. This is due to the fact that the power subtraction algorithm as well as the log-MMSE algorithm does not treat the two types of distortions (speech distortion and noise estimation distortion) differently.

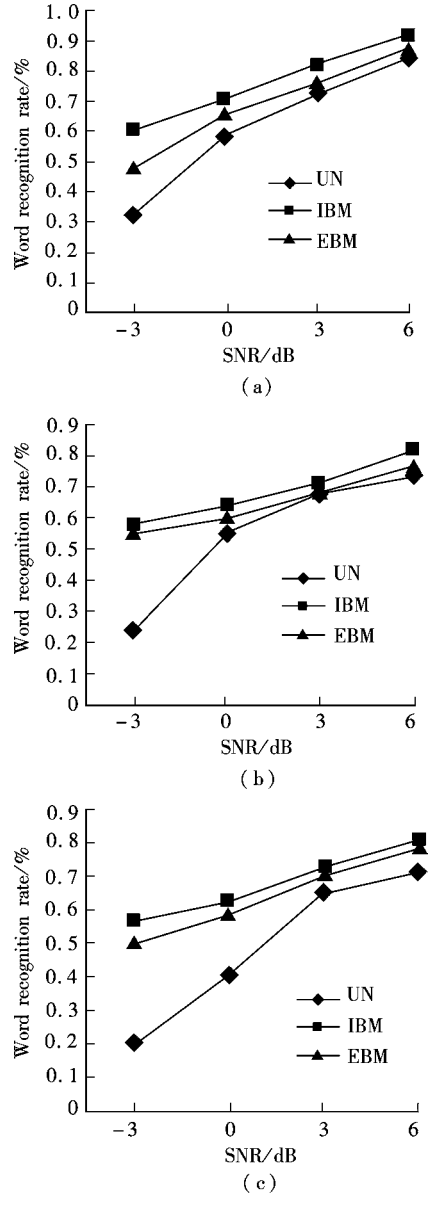

Fig. 4 Word identification rates of UN, IBM and EBM with SNRs of −3, 0, 3 and 6 dB in different noise environments. (a) Gaussian noise environment; (b) Car noise environment; (c) Babble noise environment

Tab. 1 Word identification rates of the proposed algorithm, power subtraction and log-MMSE

Noise	SNR/dB	EBM	PS	Log-MMSE
GWN	−3	0.47	0.30	0.38
	0	0.59	0.45	0.51
	3	0.73	0.60	0.62
	6	0.88	0.76	0.74
F16	−3	0.56	0.27	0.35
	0	0.64	0.42	0.50
	3	0.72	0.57	0.61
	6	0.85	0.74	0.73
Babble	−3	0.4	0.20	0.30
	0	0.56	0.33	0.42
	3	0.67	0.51	0.54
	6	0.78	0.62	0.63

4 Conclusion

A new algorithm based on the binary mask for whispered speech enhancement is proposed to improve whispered speech intelligibility in adverse noisy environments. The pursued approach focuses on the reliable classification of the T-F unit of the noisy whisper using a supervised machine learning methodology. Each T-F unit is retained or discarded according to the binary mask estimated by the trained SVM classifier, where a noise robust feature is used. The proposed algorithm is evaluated using objective and subjective evaluation, i. e., the STOI value and the listening test with normal-hearing listeners, respectively. Experimental results indicate that the intelligibility of whispered speech processed by the proposed algorithm is substantially higher than that of the unprocessed whispered speech, as well as that of the enhanced whisper using the power subtraction algorithm and the log-MMSE algorithm. Intelligibility is improved by suppressing the background noise without distorting the underlying target speech signal. We attribute this to the accurate classification of the T-F units into the target- and masker-dominated T-F units, and the subsequently reliable estimation of the binary mask.

References

[1] Tartter V C. What's in a whisper? [J]. *The Journal of the Acoustical Society of America,* 1989, **86**(5): 1678 - 1683.

[2] Ito T, Takeda K, Takura F. Analysis and recognition of whispered speech [J]. *Speech Communication*, 2005, **45**(2): 139 - 152.

[3] McAulay R, Malpass M. Speech enhancement using a soft-decision noise suppression filter [J]. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980, **28**(2): 137 - 145.

[4] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator [J]. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1985, **33**(2): 443 - 445.

[5] Loizou P C, Kim G. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(1): 47 - 56.

[6] Cooke M, Ellis D P W. The auditory organization of speech and other sources in listeners and computational models [J]. *Speech Communication*, 2001, **35**(3/4): 141 -177.

[7] Bregman A S. *Auditory scene analysis: the perceptual organization of sound* [M]. Cambridge: The MIT Press, 1994.

[8] Wang D L, Kjems U, Pedersen M S, et al. Speech intelligibility in background noise with ideal binary time-frequency masking [J]. *The Journal of the Acoustical Society of America*, 2009, **125**(4): 2336 - 2347.

[9] Li N, Loizou P C. Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction [J]. *The Journal of the Acoustical Society of America*, 2008, **123**(3): 1673 - 1682.

[10] Varga A, Steeneken H. Assessment for automatic speech recognition: Ⅱ. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems [J]. *Speech Communication*, 1993, **12**(3): 247 -251.

[11] Taal C, Hendriks R, Heusdens R, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(7): 2125 - 2136.

[12] Hu Y, Loizou P C. A comparative intelligibility study of single-microphone noise reduction algorithms [J]. *The Journal of the Acoustical Society of America*, 2007, **122**(3): 1777 - 1786.

基于噪声鲁棒性特征和 SVM 的耳语音可懂度增强

周 健[1,2] 赵 力[1] 梁瑞宇[1] 方贤勇[2]

([1] 东南大学水声信号处理教育部重点实验室,南京 210096)

([2] 安徽大学智能计算与信号处理教育部重点实验室,合肥 230601)

摘要:提出了一种基于机器学习的耳语音可懂度增强方法. 该方法利用已经训练好的 2 类支持向量机来估计一个二元时频掩蔽值,进而合成增强后的耳语音. 输入支持向量机的特征向量 GFCCs 是基于听觉外周模型进行提取的,具有噪声鲁棒特性. 在增强仿真实验中,将该算法同传统语音增强算法进行语音可懂度增强性能比较. 客观评价和主观听力实验结果均表明,所提出的方法能有效提高含噪耳语音的听觉可懂度;相比谱减法和 log-MMSE 方法在低信噪比时无法提高语音可懂度,该方法在低信噪比时仍可有效提高含噪耳语音的听觉可懂度. 此外,含噪耳语音通过所提出的方法进行增强后,其可懂度比未增强时明显提高.

关键词:耳语音;可懂度增强;噪声鲁棒性特征;机器学习

中图分类号:TN912. 35