Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **6**, November **2018**

Two-stage Vocal Effort Detection Based on Spectral Information Entropy for Robust Speech Recognition

Hao Chao*, Liang Dong, Yongli Liu

School of Computer Science and Technology
Henan Polytechnic University
Jiaozuo 454000, China
*chaohao1981@163.com

Received December, 2017; revised April, 2018

Abstract. *The features commonly used in vocal effort detection, such as global spectrum features and MFCC, have a strong ability to distinguish whisper mode, but have a bad performance when detecting the other VE modes. So in this paper a two-stage detection framework based on spectral information entropy features is proposed for the classification of vocal effort levels in robust speech recognition. Firstly, global spectrum features are used to judge whether the current speech signal belongs to the whisper level. Then, for the speech signal which does not belong to the whisper level, frame-level spectral information entropy features are acquired from vowel segments, and vocal effort level of the speech signal is determined by vowel template matching method. Finally, proposed vocal effort detection method is integrated into the robust speech recognition system based on multiple modes framework. Experiments conducted on isolated words test set show that accompanied by a slight increase for whisper level, significantly improvement of recognition accuracy for the remaining four vocal effort levels can be achieved, and the multiple modes framework is able to effectively deal with the mismatch of the training environment and test environment.*

Keywords: Robust speech recognition, Vocal effort, Spectral information entropy, Gaussian mixture model, Template matching

1. Introduction.

Vocal effort(VE) was characterized as the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance[1].Generally, there are five different vocal effort levels(whispered, soft, normal, loud, and shouted). Changes in vocal effort result in a fundamental change in speech production and then cause the change of acoustic characteristics, which will reduce the accuracy of speech recognition system [2]. Therefore, accurate VE detection can enlarge the application range of speech recognition technology, and will promote the practicability of speech recognition. In addition, it also has a positive effect on speaker recognition and speech synthesis [3,4].

In comparison with other vocal effort levels, the whispered speech is obviously different in speech production mechanism and acoustic characteristic because the vocal cords are almost not vibrating.Therefore, as a typical representative of VE, related studies of whisper have been conducted since the 1960s, and the accuracy of whisper detection is satisfactory. In literature [5], the average energy ratio between high energy segment and low energy segment of low-frequency band is acquired, and the ratio is used as a basis for the judgment of a whisper speech or a normal voice. Zhang and Hansen proposed a detection method of vocal effort change points [6]. In literature [7], a whisper detection algorithm is proposed using features obtained from waveform energy and wave period. In addition, M Sarria-Paja and TH Falk propose an accurate whispered speech detection method at signal-to-noise ratios, which uses auditory-inspired modulation spectral-based features to separate speech from environment-based components[8].

For the remaining four vocal effort levels, there are no significant differences in the way of pronunciation, and no significant changes have been reflected in the spectrum. For the two adjacent vocal effort levels, it is even more so. Therefore, just a few studies collectively consider detection of all five speech levels, and only limited performances are provided.A first attempt is presented by using Gaussian mixture models (GMM) which are trained by global spectrum features which are composed of sound intensity level, sentence duration, frame energy distribution and spectral tilt[9,10]. These features are very effective in detecting whisper, but only limited results have been achieved when identifying the remaining four VE levels. In the literature [2], a VE classification method using Support Vector Machine (SVM) is proposed based on the Mel-frequency cepstral coefficients (MFCC), and better results are achieved. Nevertheless, MFCC is proposed for speech recognition, so this feature mainly reflects acoustic properties caused by different pronunciation instead of vocal effort change.

In order to further improve the detection accuracy of all five VE levels, this paper presents a two-stage detection method. In the method, GMMs trained by global spectrum features are used to judge whether the current speech signal belongs to the whisper level. For the speech signal which does not belong to the whisper level, spectral information entropy (SIE) which showed a strong ability to distinguish the remaining four VE levels is acquired from vowels, and vocal effort level of the speech signal is determined by vowel template matching method which using the spectral information entropy.

This paper is organized as follows: In Section 2, the introduction of spectral information entropy is given. Section 3 introduces the proposed two-stage VE detection method and the multiple model framework approach for robust speech recognition. The performance of the proposed method is reported in Section 4. The last Section 5 briefly concludes the work.

2. **Spectral Information Entropy.**

2.1. **Feature extraction.** For each frame, the spectrum obtained from FFT can be viewed as a vector of coefficients in an orthonormal basis. Hence, the probability density function (pdf) can be estimated by the normalization over all frequency components. The spectral information entropy can be obtained from this estimated pdf.

Each frame is evenly divided into 6 sub-bands. The SIE of each sub-band is calculated to form a 6-Dimension SIE feature for each frame. In fact, the 6 dimensions of the feature are the spectral information entropy of the 6 sub-bands evenly divided over the frequency range 0-4000 Hz, respectively.The six bands and their range of frequency domain are shown in Table 1.

For each sub-band, the spectral information entropy can be obtained as follows:

Assuming X is the power spectrum of a sub-band of the current speech frame, X(k) represents the k-th frequency component in the sub-band, and k varies from k_1 to k_M. Then the portion of frequency content in the k-th frequency component versus the entire sub-band is written as,

$$p(k) = \frac{|X(k)|^2}{\sum_{j=k_1}^{k_M} |X(j)|^2} \ , \ k = k_1, ..., k_M \tag{1}$$

Table 1. Six bands and their range of frequency domain

Sub-band	Frequency range (kHz)
1	0.0-0.8
2	0.6-1.5
3	1.2-2.0
4	1.8-2.6
5	2.4-3.2
6	3.0-4.0

Since $\sum_{k=k_1}^{k_M} p(k) = 1$, p(k) has the property of probability. The spectral information entropy for the sub-band can be calculated as,

$$H = -\sum_{k=k_1}^{k_M} p(k) \cdot \log\ p(k) \tag{2}$$

Using the power spectrum of each frame, the above calculation is performed for each of the 6 sub-bands in Table 1, so that the 6-D SIE over the frequency domain is obtained for each frame.

2.2. **Salient information analysis of SIE.** From the perspective of speech perception, speech signals are composed of vowels, consonants and silent segments. Obviously, silent segment does not contain salient information regarding the VE level. So we only need to know which contains more salient information between vowel and consonant.

In order to facilitate the analysis, it can be assumed that the speech signal, which shows greater spectrum change when VE level changes, contain more salient information regarding the VE level. For this purpose, a Euclidean distance-based cepstral distance measure is used:

$$D_C = \sqrt{\sum_{k=1}^{N} \left(c_p^{VE_i}(k) - c_p^{VE_j}(k)\right)^2} \tag{3}$$

where D_c represents the spectral distance of two speech signals which are produced when the same phoneme p is pronounced under VE level i and VE level j respectively, SIE is used as spectral features. N is the number of dimensions of spectral features. $c_p^{VE_i}$ represents the mean vector of the spectral feature sequence of the speech signal produced when the same phoneme p is pronounced under VE level i, and $c_p^{VE_i}(k)$ is the k-th component of the vector $c_p^{VE_i}$. An average distance between all pairs of VE levels for a given phoneme was then computed. After normalization, the obtained average distances for all phonemes are documented in Fig.1 (sorted in descending order). The highest average distances were obtained for the set of 5 vowels (/a/, /e/, /o/, /i/, /u/) and consonants/j/, /g/, and /y/. These consonants appear less frequently in words, hence the vowels are the best candidates for VE classification.

The average cepstral distances using MFCC features are also acquired in this paper, and the average cepstral distances are compared with the average cepstral distances of SIE. After the analysis above, we only compare the five Chinese vowels: a e o i u. As shown in Fig.2, when using SIE features, the average spectral distance of each vowel is higher than that of the MFCC feature. This seems to indicate that SIE contain more salient information regarding the VE level.

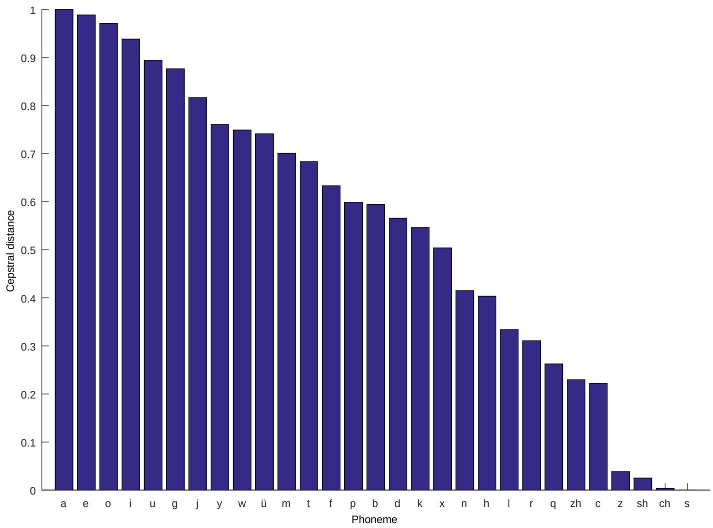

Figure 1. Sorted average cepstral distances among the 5 VE levels for all phonemes

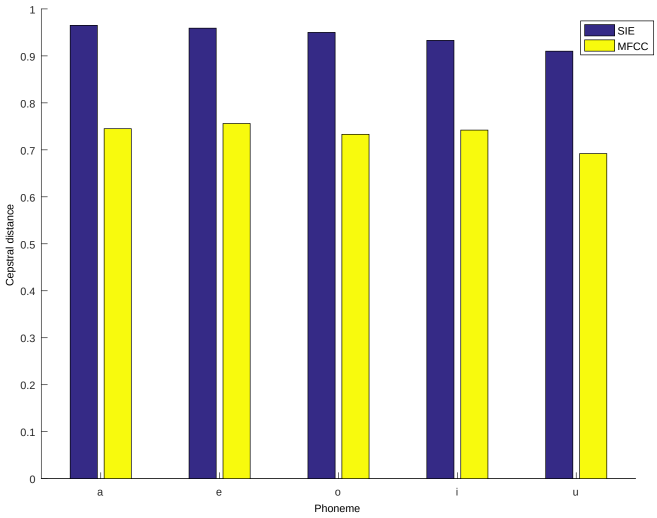

Figure 2. Comparison of average cepstral distances between MFCC and SIE for the five vowels

It is important to keep in mind that the speech samples grouped into the individual VE levels are not results of some artificial signal classification but a genuine representation of what the speakers considered to be whispering, soft speech, normal speech, etc. The histograms therefore establish a connection between a subjective quantity (the VE level) and a measurable physical quantity.

3. Two-stage VE Detection and Multiple Model Framework for Speech Recognition.

3.1. Two-stage VE detection. The proposed VE detection is shown in Fig. 3, including the two-stage identification process.

The purpose of the first-stage identification is to judge whether the current speech signal is a whispered speech. In the stage, sound intensity level, syllable duration, frame energy distribution and spectral tilt are used to train two GMMs. One GMM corresponds to the whisper level, and the other GMM corresponds to the other four levels. Each GMM has 32 Gaussian mixtures, and each Gaussian mixture employs diagonal covariance matrix.

If the current speech signal is judged to be a whispered speech, the result is the final recognition result. Otherwise, the second stage identification is needed to identify specific VE level. In the second stage, speech segments corresponding to vowels are acquired from the speech signal by manual segmentation and frame-level spectral information entropy is extracted from the vowels. Then, the vowel template matching method is proposed to determine VE level of the speech signal. It should be noted, we only know that the detected speech segment is a vowel, but do not know what is the specific vowel.

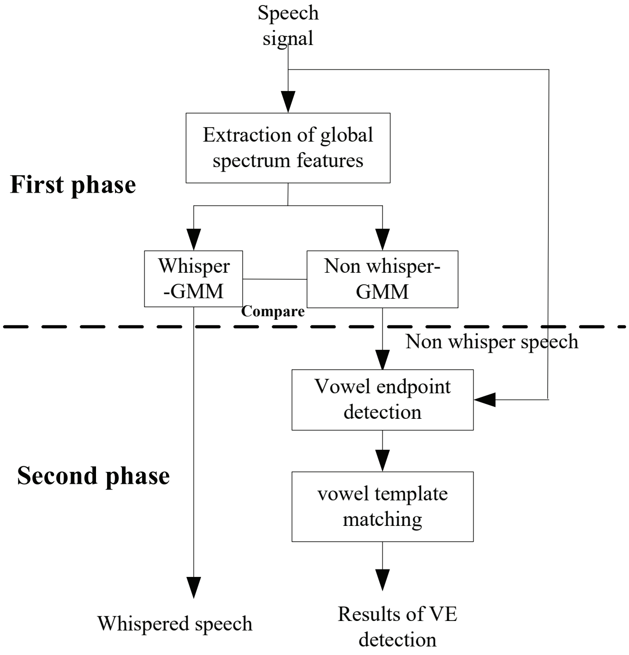

Figure 3. Two-stage detection of vocal effort

3.2. **Vowel template matching.** Because it is not sure what kind of vowel corresponds to the detected speech segment, for each VE level (except for whisper level), a corresponding standard vowel template set is set up. Each standard vowel template set contains 5 simple vowel templates (a,o,e,i,u). A vowel template is defined as a standard pronunciation unit of the vowel under the corresponding VE level.

Firstly, for a detected vowel segment in the current speech signal S, the cepstral distance between the detected vowel segment and each VE level is measured:

$$D_{(v,VE_j)} = \min_p \sqrt{\sum_{k=1}^{N} (c_v(k) - c_p^{VE_j}(k))^2} \tag{4}$$

where $D_{(v,VE_j)}$ represents the cepstral distance between the detected vowel segment v and the specific VE level j, and c_v is the mean vector of SIE sequence belonging to v, and $c_v(k)$ is the k-th component of c_v. VE_j is the specific VE level j, and p is a vowel template belonging to the standard vowel template set of VE_j. $c_p^{VE_j}$ is the mean vector of SIE sequence belonging to p. $c_p^{VE_j}(k)$ is the k-th coefficient of $c_p^{VE_j}$. For any vowel template, five candidate of the vowel template are acquired from the train set by manual segmentation.For each candidate, the mean vector of SIE sequence is calculated. Then, average of the five candidates mean vectors is calculated as the mean vector of SIE sequence belonging to this vowel template.

Secondly, the total cepstral distance between the current speech signal S and each VE level is obtained:

$$D_{(S,VE_j)} = \sum_{v \in T} D_{(v,VE_j)} \tag{5}$$

where $D_{(S,VE_j)}$ represents the total cepstral distance between the current speech signal S and the specific VE level j, and T is the set of the detected vowel segments in the speech signal S.

Finally, VE level of the current speech signal S is decided:

$$\tilde{VE} = \arg\min_{VE_j} D(S, VE_j) \tag{6}$$

3.3. **Multiple-Model Framework.** The multiple-model framework (MMF) was previously employed for noise-robust speech recognition [11]. It was shown that acoustic models perform better if only one type of noisy speech is covered as opposed to the multi-style trained universal hidden Markov models (HMMs). Consequently, several model sets are needed to cover the whole range of possible noise conditions. An integral part of the MMF is a model selector, i.e. a classifier determining the best HMM set for the current conditions.

We propose to apply the MMF approach to achieve speech recognition that is immune to VE changes, as shown in Fig.4. The five VE levels (whispering, soft, normal, loud, and shouting) compose the MMF classes among which a specialized VE classifier chooses the one that best matches the speakers actual speaking level. Each class is assigned a dedicated acoustic model set trained by the specifics of speech in the given level, and HMM is used as the acoustic model.

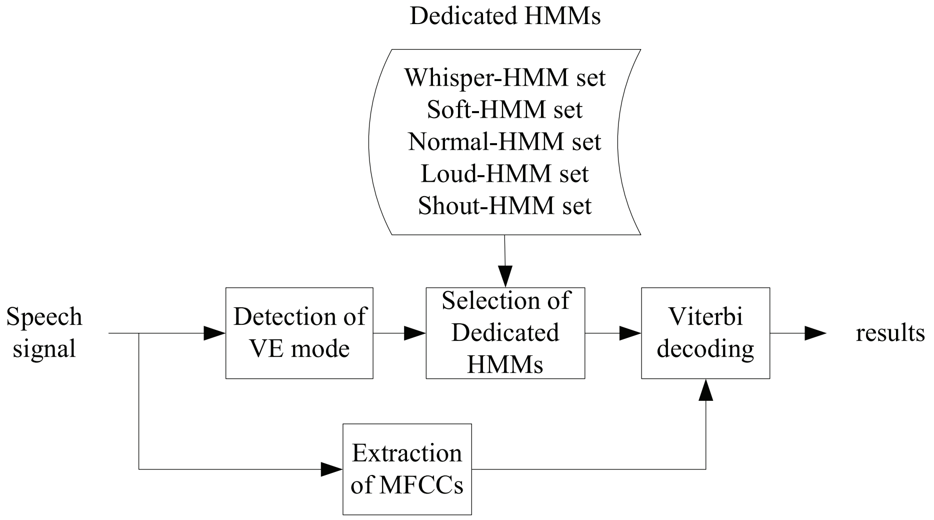

FIGURE 4. Block diagram of the multiple-model framework

4. Experimental Classification Results and Analysis.

4.1. Speech corpora.

The data corpus applied in experiments consists of 25000 Mandarin isolated digits(0-9). 20 male speakers are employed for train set and test set. In the train set, each VE level contains 4000 digits, and each speaker record 20 times the digits (0-9). In the test set, each VE level contains 1000 digits, and each speaker record 5 times the digits (0-9). The data corpus is recorded in the laboratory environment, and is stored using the 16 kHz sampling rate and 16-bit resolution.

Each dedicated acoustic model set has 10 HMMs, corresponding to the 10 Arabia figures respectively. The structure of HMM is left to right with 5 states, 3 emitting distributions and no state skipping. Each emitting distribution is modeled by 16 Gaussian mixtures. The HMMs is developed by HTK V3.2.1 [12].

4.2. VE level classification.

It is important for the proposed MMF-based speech recognition to achieve a reliable detection of the current VE level determined from the speech signal.

Firstly, we employ the global spectrum features proposed in literature 8 to construct a VE detection baseline system to detect all VE levels. Classification model adopts GMM, and each VE level corresponds to a GMM which is trained by the global spectrum features acquired from speech of the current VE level. The detailed results are shown in Table2.

Table 2. VE detection results by using global spectrum features

Actual	Detection result(%)				
VE level	Whisper	Soft	Normal	Loud	Shouted
Whisper	96.2	3.8	0	0	0
Soft	2.8	69.5	27.7	1.2	0
Normal	0	22.7	60.5	16.8	0
Loud	0	0	21.1	65.7	13.2
Shouted	0	0	0	28.8	71.2

As can be seen from Table 2, good performance has been achieved for whisper level detection, and the result reveals that the four global spectrum features contain salient information regarding whisper level, just due to the significant spectrum difference between whisper speech and speech of other VE levels. However, the recognition accuracy has been greatly decreased for the other four VE levels, and detection errors mainly occur in the adjacent VE levels. This shows that the four global spectrum features dont contain adequate discriminative information to identify the remaining four VE levels, in particular the adjacent VE levels. The average recognition accuracy of the five VE levels reaches 72.6%.

Then, the proposed two-stage method is used to identify the VE level of the test set, and the recognition results are shown in Table 3. In addition, as an alternative to SIE, 12 dimensions MFCC are also employed based on the same two-stage detection framework, and the recognition results are shown in Table 4.

As can be seen from Table 3, the proposed method achieves 96.2% whisper recognition accuracy, which is the same as that of the baseline system. This is mainly because the global spectrum features are also used to detect whisper level in the first-stage identification process of the proposed method. Owing to the same reason, whisper recognition accuracy in Table 4 is also 96.2%.

The recognition accuracies of soft level, normal level, loud level and shouted level in Table 3 are 82.4%, 78.2%, 82.3% and 84.9% respectively. Compared with the baseline

Table 3. Two-stage VE detection results by using SIE

Actual	Detection result(%)				
VE level	Whisper	Soft	Normal	Loud	Shouted
Whisper	96.2	3.8	0	0	0
Soft	3.3	82.4	14.3	0	0
Normal	0/0	11.7	78.2	10.1	0
Loud	0/0	0/0	11.8	82.3	5.9
Shouted	0/0	0/0	0/0	15.1	84.9

Table 4. Two-stage VE detection results by using MFCC

Actual	Detection result(%)				
VE level	Whisper	Soft	Normal	Loud	Shouted
Whisper	96.2	3.8	0	0	0
Soft	4.1	78.2	17.7	0/0	0/0
Normal	0/0	16.9	71.5	11.6	0/0
Loud	0/0	0/0	15.5	74.6	9.9
Shouted	0/0	0/0	0/0	19.3	80.7

system (shown in Table2), the proposed method has overwhelming advantages in identifying these four VE levels. This is mainly because vowels have more salient information regarding the VE level than the global spectrum features and consonants, and SIE feature which reflecting the energy distribution of each frequency band can extract the salient information effectively. In addition, it is difficult to detect specific vowels in speech signals accurately. Therefore, we can only obtain the vowel segments in the speech signal. This means that we only know that the detected speech segment is a vowel, but do not know what is the specific vowel. In view of this situation, the method of vowel template matching is used in the proposed two-stage framework, which can exploit the salient information contained in vowel segments to determine the VE level of the current speech without knowing the specific vowels. And the average recognition accuracy of all VE levels in Table 3 reaches 84.8%.

The recognition accuracies of soft level, normal level, loud level and shouted level in Table 4 are 78.2%, 71.5%, 74.6% and 80.7% respectively. Although these results are better than the baseline recognition results, they still fail to keep pace with the results of the proposed methods in Table 3. The results also show that SIE is able to extract the salient information of vowel segments better than MFCC. And the average recognition accuracy of all VE levels in Table 4 reaches 80.3%.

4.3. **Robust speech recognition.** The isolated word recognition experiment is performed after VE level classification. In order to evaluate the effect of VE change on the performance of speech recognition, each dedicated acoustic model set is employed to perform decoding of speech under all VE levels, and the results are shown in Table 6. The most left column in Table 5 represents the dedicated HMM sets, and recognition results of speech under each VE level by using current dedicated HMM set are shown on the right side of the table.

For each VE level, good word recognition performance can be achieved when the corresponding dedicated HMM set is employed. However, the recognition accuracy will be greatly reduced when using other HMM set, due to the mismatch of the training environment and test environment. Especially for VE levels with large scale difference, for

Table 5. Word error rate (WER) of dedicated HMMs for each VE level

Dedicated	WER of speech under each VE level(%)				
HMM set	Whisper	Soft	Normal	Loud	Shouted
Whisper	0.51	39.56	30.62	46.69	59.27
Soft	57.74	2.98	1.88	31.83	42.08
Normal	56.24	13.48	0.65	19.32	37.84
Loud	51.75	36.31	17.29	1.12	21.13
Shouted	50.23	42.85	37.26	19.88	2.91

example whisper and shouted voice, the recognition accuracy is lower when identifying each other speech by using own dedicated HMM set.

Finally, isolated word recognition experiment is carried out on the test set according to the proposed MMF recognizer, and the results are shown in Table 6.

Table 6. WER of MMF recognizer

WER(%)				
Whisper	Soft	Normal	Loud	Shouted
0.94	5.36	2.78	3.95	4.32

Compared with recognition results using the dedicated HMM set of normal VE level(shown in the third row in Table 6) , recognition accuracy of all VE levels except normal level are greatly improved. And speech recognition performance of normal VE is decreased, which is caused by the improper selection of acoustic model set caused by the error of VE detection.

5. **Conclusion.** In this paper, after analyzing the sensitivity of MFCC and SIE to the change of VE level, we proposed a two-stage detection framework by using frame-based the spectral information entropy. Average accuracy of the baseline system achieved 72.6%, while the proposed VE classifier reached 84.8%. This yields a 12.2% absolute increment.

Moreover, we also analyzed the impact of varied vocal effort level on the performance of automatic speech recognition in all speech levels, ranging from whispering to shouting. An isolated-word speech recognizer utilizing whole-word hidden Markov modes with Gaussian mixture output distributions was used in the experiments.

The main contributions of this article are as follows: 1) spectral information entropy features for VE detection are proposed; 2) realizing the fusion of global spectrum features and SIE features by the two-stage VE detection; 3) the multiple-model framework is introduced into the VE related robust speech recognition.

The future research will be focused on a more precise detection of VE level considering real-world situations (i.e. including additive noise.). Reliable vowel endpoint detection is very important for the proposed detection framework, and it is also our future research priorities.

Acknowledgment. This paper is funded by the China National Nature Science Foundation (No.61502150, No.61300124, No.61403128), Foundation for University Key Teacher by Henan Province under Grant number 2015GGJS-068 and the Fundamental Research Funds for the Universities of Henan Province.

REFERENCES

[1] H. Traunmller , A. Eriksson, Acoustic effects of variation in vocal effort by men, women, and children, *Journal of the Acoustical Society of America*, vol. 107, no.6, pp.3438-3451, June, 2000.

[2] P. Zelinka , M. Sigmund , and J. Schimmel, Impact of vocal effort variability on automatic speech recognition, *Speech Communication*, vol.54, no.6, pp. 732-742, 2012.

[3] R. Saeidi, P. Alku, and T. Backstrom, Feature extraction using power-Law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch, *IEEE Transactions on Audio, Speech, and Language Processing*, 24(1) : 42-53, 2016.

[4] T. Raitio , A. Suni , J. Pohjalainen,et al, Analysis and synthesis of shouted speech *INTERSPEECH*, pp. 1544-1548, 2013.

[5] S. J. Wenndt , E. J. Cupples , and R. M. Floyd, A study on the classification of whispered and normally phonated speech, *7th International Conference on Spoken Language Processing*, pp. 649-652, 2002.

[6] C. Zhang, H. L. John, and Hansen, Advancements in whisper-island detection within normally phonated audio streams, *10th Annual Conference of the International Speech Communication Association*,pp.860-863, 2009.

[7] M. A. Carlin, B. Y. Smolenski, and S. J. Wenndt, Unsupervised detection of whispered speech in the presence of normal phonation, *9th International Conference on Spoken Language Processing*, pp. 685-688, September 17-21,2006.

[8] M. Sarria-Paja, T. H. Falk, Whispered speech detection in noise using auditory-inspired modulation spectrum features, *IEEE Signal Processing Letters*, vol.20,no.8, pp.783-786, 2013.

[9] P. Zelinka , M. Sigmund, Automatic vocal effort detection for reliable speech recognition, *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pp.349-354, 2010.

[10] C. Zhang, H. L. John, and Hansen, Analysis and classification of speech mode: whispered through shouted, *8th Annual Conference of the International Speech Communication Association*, pp.22892292, 2007.

[11] H. Xu , Z. H. Tan , P. Dalsgaard , and B. Lindberg, Robust speech recognition based on noise and SNR classification - a multiple-model framework, *9th European Conference on Speech Communication and Technology*, pp. 977980, 2005.

[12] S. Young, G. Evermann, and M. Gales, et al, Hidden markov model toolkit[EB/OL], *Cambridge: Cambridge University Engineering Department*, http://htk.eng.cam.ac.uk/docs/docs.shtml, 2000.