An Audio-Visual Whisper Database in Chinese

Jian Zhou¹², Yuting Hu², Hailun Lian², Cong Pang², Huabin Wang¹², Liang Tao¹

¹Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei, AnHui, 230039, China
²School of Computer Science and Technology, Anhui University, Hefei, AnHui, 230601, China

*Corresponding author’s e-mail: jzhou@ahu.edu.cn

Abstract. Converting whisper to normal vocalized speech has been a hot research topic in speech signal processing area. A complete and large scale whisper database is a major basis for this task. In this paper, we propose a multimodal whisper database in Chinese mandarin. A total of 103 syllables and 100 sentences were carefully selected. 5 male and 5 female participants pronounced the syllables and sentences in whisper and normal styles respectively, resulting in 4096 parallel speech utterances and 263,849 frames of voicing face and lip image sequences. The beginning and ending sample point of each syllable were labeled both for speech signal and voicing face video. The lip region of interest were also extracted and provided in the proposed database. Experiments in various speech conversion tasks in different speech database show the effectiveness of the proposed multimodal whisper speech database.

1. Introduction

Whisper refers to the low-energy pronunciation with no vocal-cord vibrating. It is a special and important communication style between humans. In places such as library or meeting room, people always adopt whisper due to the prohibition of loud noise and the protection of communication privacy. Whisper is also the only way to communication for laryngeal cancer patients who have suffering from aphonia and cannot pronounce normally.

Due to the property of low energy and pitch absence, whispered speech has gain a lot of attention in recent years. The existed research can be divided into three categories: whispered speech recognition [1-2], whispered speech emotion recognition [3-4] and whisper to normal speech conversion [5-7].

Whisper to normal speech conversion is an emerging research topic in speech signal processing because of its potential application to enable people, who have injured their vocal cords, to communicate using their own voice without help of medical equipment such as an electric artificial larynx. Currently, spectral mapping techniques using Gaussian mixture models (GMM)[8] or Deep Neural Networks (DNN)[9] borrowed from voice conversion have been applied to transform whisper spectral features to normally phonated audible speech.

However, the existing conversion models are trained ideally with corpus obtained in quiet environment. Its performance decreases obviously in noisy environment. This is inconsistent with the auditory perception abilities of human beings. In fact, speech perception is essentially a multi-channel processing process. Visual information plays a pivotal role in noisy environment. For example, [10-11] have shown that visual modality enhances the performance of speech processing compared with its counterpart that uses audio modality alone. There is an obvious complementary relationship between
visual information and auditory information, and visual information can be seen as an auxiliary in speech processing [12].

Multimodal whisper to normal speech conversion aims to obtain high quality phonated speech from whisper contaminated by various types of noise. Similar to lip-reading, the face or lip movement information is adopted as a key complementary feature in multimodal whisper to normal speech conversion process based on audiovisual speech corpus. However, most existing audiovisual corpus is recorded only by phonated speech [13-14]. There are only two public audiovisual corpus database of whisper. The first one was built in Japanese in 2002 [15]. The other audiovisual whisper corpus database was recorded in English in 2013 [16].

As far as we know, there is no public audiovisual whisper corpus in Chinese. To address the issue of whisper to normal speech conversion in Chinese, we build an audiovisual whisper corpus database in Chinese. 5 male and 5 female professional recorders are participated in the recording process in a quiet lab environment. A total of 4060 corpus along with 263849 face movement images are obtained. In order to facilitate researchers to carry out other whisper related research, the region of interest of face and lip are further extracted, and 106 key points of face and lip are labeled and stored in the database. The recorded speech signals are also post-processed where the beginning and ending of each syllable are extracted and stored in the database.

Table 1. Audio-visual whisper database configuration in [15] and [16].

|            | [15]                          | [16]                          |
|------------|-------------------------------|-------------------------------|
| Speakers   | 8 males, 3 females            | 62 males, 49 females          |
| Corpus     | 129 sentences, 11 isolated digits | 60 sentences                  |
| Number of pronunciations | >=3960                      | >=6000                        |
| Pixel      | 1440*1080                     | 720*480                       |
| Fps        | 30fps                         |                               |
| Sampling frequency | 48KHZ                      | 16KHZ                         |
| Quantization accuracy | 16bit                      | 16bit                         |

2. The proposed multimodal whispered speech database in Chinese

In this section, we give the construction detail of the multimodal whispered speech corpus database in Chinese. Generally, the participated speakers utter each specified corpus both in phonation and whisper style. The parallel whisper and normal speech utterances, along with the lip movement sequences are recorded and preprocessed. The detailed configuration is illustrated in table 2.

Table 2. The multimodal whispered speech database in Chinese.

|                   | 5 males, 5 males                        |
|-------------------|----------------------------------------|
| Speaker           |                                        |
| Sentences and words| 100 sentences, 103 isolated syllable    |
| Number of pronunciations | 4060                              |
| Pixels            | 1920*1080                              |
| fps               | 30fps                                  |
| Total number of frames | 263,849                       |
| Sampling frequency | 44.1KHZ                             |
| Quantization accuracy | 16bits                         |

2.1 Corpus selection

The AVWD is focused on whispered speech analysis in Chinese and is used mainly for whisper to normal conversion. To this end, the articulation difference between whisper and normal speech in Chinese is a major factor for corpus selection. When the Mandarin is pronounced, only the vowel and
the voiced consonant (including nasal consonant) comprise pitch, which is, however, absent for voiceless consonant, making its pronunciation similar to whispered speech.

In Mandarin, all of the final (single and vowels) are vowels, and the initials can be categorized into six types, i.e., clear sound, squeak, stop sound, side, nasal, and zero initials. Among them, the nasal, side and zero initials are voiced consonants, while others are voiceless consonants [13]. According to the mouth shape of the pronunciation, the Mandarin syllable can be sub-divide into 46 categories [17]. In this paper, all of the syllable initials are voiced consonant. A total of 94 syllables of 27 types are selected according to the pronunciation differences of whisper and normal speech. In addition, we add syllables of digit 0 to 9. A total of 100 consecutive short sentences were selected as the multimodal corpus of whisper from the Chinese corpus developed by the Chinese Academy of Sciences Automation.

2.2 Data collection and post-processing
Visual-audio speech data is collected in a quiet environment. The video capture device with a resolution of 1920*1080 and a frame rate of 30fps is used for video sequences recording. Each speaker pronounces the aforementioned speech corpus in Chinese and takes a rest every 10 minutes. Once the data collection stage is finished, we have a further process for the recorded video data. A total of 106 key feature points of the speaker face are labeled. The region of lip and face area of each speaker are also extracted. The lip area along with the facial area and the key feature points from a video image sequence are shown in figure 1.

For each of the recorded speech corpus, the start and end sampling points for each syllable are labeled as shown in figure 2. Finally, the beginning and ending frame of video for each phonated syllable are derived and aligned to labeled speech data.
3. Experiments and discussion
To test the validity of the proposed multimodal speech database, three types of experiments were conducted. The first one is male to female voice conversion just based on speech, and the second experiment is about whisper to normal speech conversion. The last experiment is multimodal speech conversion with the presence of noise. We adopt MFCC as acoustic feature in this paper. GMM and DNN were used for the first and second experiments [18-19]. For comparison, the male to female conversion based on multimodal was also conducted based on AVDCNN model which was used for speech enhancement in [11] and DNN model.

In this paper, the ABX [20] was used as the subjective evaluation which was formalized as,

$$ABX = \sum_{n=1}^{M} \sum_{m=1}^{N} P_{n,m}$$

(1)

Where \(M\) and \(N\) represent the number of tester and test voice, respectively. \(P_{n,m}\) is the tendency of the target determined by \(m\)-th tester for \(n\)-th converted speech. The range is 0 to 1, higher score means higher similarity between converted and referenced speech.

For objective evaluation, we adopt the Cepstral Distortion (CD) defined as follows.

$$CD = \frac{1}{10 \log 10} \sqrt{\sum_{d=1}^{D} (C_{d} - C_{d} ')^2}$$

(2)

Where \(C_{d}\) and \(C_{d}'\) are, respectively, the \(d\)-th Mel-cepstral of the converted speech and the referenced speech. \(D\) is the dimension of Mel-cepstral which was set to 26 in our experiments. The higher CD value denotes greater difference between the converted and referenced speech.

A total of 100 aligned parallel utterances were taken from AVWD and CASIA [21] respectively, 90 of which were randomly selected for training and the remainder for testing data. Both the Gaussian Mixed Model (GMM) and Deep Neural Network model (DNN) were used for speech conversion. Table 4 gives the scores of CD and ABX of converted speech with each conversion model in AVWD and CASIA databases respectively.

Table 3. CDs and ABXs of male to female speech conversion.

| Method | AVWD | CASIA | AVWD | CASIA |
|--------|------|-------|------|-------|
| Source | CD   | ABX   | CD   | ABX   |
| GMM    | 6.7173 | --   | 7.1231 | --   |
| DNN    | 5.5533 | 100% | 6.6568 | 100% |
|       | 4.9371 | 100% | 5.4266 | 100% |

Table 4. CDs and ABXs value of whisper to normal speech conversion based on different databases.

| Method | AVWD | TIMIT | AVWD | TIMIT |
|--------|------|-------|------|-------|
| Source | CD   | ABX   | CD   | ABX   |
| GMM    | 8.2729 | --   | 8.0098 | --   |
| DNN    | 6.4529 | 100% | 6.0168 | 100% |
|       | 6.3808 | 100% | 5.6274 | 100% |

From table 3, one can find that the mean CD scores are reduced by 1.164 and 0.4663 respectively on the GMM model under both databases. And the mean CD scores are reduced by 1.17802 and 1.6965 on the DNN model.

For whisper to normal speech conversion experiment, a total of 100 whisper utterances taken from TIMIT [22] were used as comparative data with AVWD in whisper to normal speech conversion based on GMM and DNN respectively. 90% of the utterances were selected as training set randomly, and the rest 10 sentences as testing data. As can be seen in table 4, the average CD based on AVWD is close to TIMIT. The ABX score are same for both whispered databases. Compared with AVWD, the TIMIT database gains a little larger mean CD improvement. We attribute this to the language difference between Chinese and English.

For male to female speech conversion, the spectrograms of the converted speech along with the target female speech and source male speech are shown in figure 3. For whisper to normal speech conversion, the spectrograms of the converted speech (right panel) along with the normal speech (left panel) and whisper (middle panel) are shown in figure 4. One can find that the spectrogram of the converted speech is very similar to the target speech in aspect of spectrogram.
Figure 3. Spectrograms of voice conversion (the top panel is based on CASIA and the bottom panel is based on AVWD. The spectrograms in left, middle and right of each panel are target female, source male and converted speech, respectively).

Figure 4. Spectrograms of whisper to normal conversion (the top panel is based on TIMIT database, the bottom panel is based on AVWD database. The spectrograms in left, middle and right of each panel are the normal, whisper and converted speech, respectively).

In order to verify the effectiveness of the face and lip feature, 100 aligned parallel utterances were taken from the AVWD for male to female speech conversion based on multimodal features. 90 utterances were randomly selected as training set, with the remaining 10 utterances used as the testing set. The male clean utterances were artificially contaminated by white noise at SNR of 20dB. Both the AVDCNN and DNN model were used for speech conversion.

Table 5. Mean CDs of Multimodal male to female conversion

| Method   | CD(mean) |
|----------|----------|
| Source   | 7.7735   |
| AVDCNN   | 5.7357   |
| DNN      | 5.9299   |
Table 5 gives CD mean scores of converted speech with different model in AVWD databases respectively. The AVDCNN use the acoustic and lip features while the DNN use only the audio feature.

From table 5, it can be seen that the mean CD value based on AVDCNN decreases by 0.1942 compared with DNN at an SNR of 20dB, showing the effectiveness of the visual data in AVWD database.

4. Conclusion
We build an audiovisual whisper database including whisper and corresponding normal speech aim to study whisper to normal speech conversion based on multimodal feature. The advantages of the proposed AVWD are three-fold. Firstly, the speech utterances were recorded in Chinese mandarin. As far as we know, this is the first open access multimodal whisper database where the parallel normal speech and the face and lip sequences were also recorded. Secondly, the database comprises various voicing style in Chinese as both the syllables and long sentence were considered. Lastly, except for the recorded origin speech signal and image sequences, post-processing were conducted. For example, 106 key points of each voicing face were annotated. The beginning and ending time of each syllable are extracted. Experimental results show effectiveness of the proposed AVWD database.

Acknowledgments
This research and development work was partly supported by Natural Science Fund Project of China under No.61301295, the Anhui Natural Science Fund Project (No.1708085MF151), and Anhui University Natural Science Research Project (KJ2018A0018).

References
[1] Chen, X., H., Zhao, X., Fan, X. (2016) Performance analysis of mandarin whispered speech recognition based on normal speech training model. In: Sixth International Conference on Information Science & Technology. Dalian. pp. 548-551.
[2] Đorđe, T.G., Slobodan, T.J. (2017) Whispered speech recognition using deep denoising autoencoder and inverse filtering. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 25, 12: 2313-2322.
[3] Deng, J., Xu, X., Zhang, Z., Frühholz, S., Schuller, B. (2016) Exploitation of phase-based features for whispered speech emotion recognition. IEEE Access. 4: 4299-4309.
[4] Sivan, D., Gopakumar, C. (2017) Emotion recognition and spoof detection from whispered speech. In: 2017 International Conference on Computing Methodologies and Communication (ICCMC). Erode. pp. 1091-1095.
[5] Nisha Meenakshi, G., Kumar Ghosh, P. (2018) Whipped speech to neutral speech conversion using bidirectional LSTMs, In: Interspeech. Graz. pp. 491-495.
[6] Pascual, S., Bonafonte, Antonio., Serra’, Joan., Gonzalez, Jose A. (2018) Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks, In: IberSPEECH 2018. Barcelona. pp. 117-121.
[7] Ferreira, A. (2016) Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information. In: 2016 International Symposium on Signal, Image, Video and Communications (ISIVC). Tunis. pp. 159-66.
[8] Toda, T., Shikano, K. (2005) NAM-to-speech conversion with Gaussian mixture models. In: Interspeech. Lisbon. pp. 1957-1960.
[9] Janke, M., Wand, M., Heistermann, T. and Schultz, T. (2014) Fundamental frequency generation for whisper-to-audible speech conversion. In: ICASSP. Florence. pp. 2579-2583.
[10] Masaka, K., Aihara, R., Takiguchi, T., Ariki, Y. (2014) Multimodal voice conversion using non-negative matrix factorization in noisy environments. In: ICASSP. Florence. pp. 1542-1546.
[11] Hou J., Wang, S., et al. (2018) Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence. 2, 2: 117-128.

[12] McGurk, H., MacDonald, J. (1976) Hearing lips and seeing voices. Nature. 264, 5588: 746-748.

[13] Ru, T., Xie, X. (2008) Design of a whispered Chinese speech database. J Tsinghua Univ (Sci & Tech). 48: 725-729.

[14] Zhao, H., Wang, Y., Wang, M., Yan, Z. (2009) Development of Chinese whispered database for speaker verification. 2009 Asia Pacific Conference on Postgraduate Research in Microelectronics & Electronics (PrimeAsia). Shanghai. pp. 197-200.

[15] Itoh, T., Takeda, K., Itakura, F. (2001) Acoustic analysis and recognition of whispered speech. IEEE Workshop on Automatic Speech Recognition and Understanding. 45, 2: 429-432.

[16] Tran, T., Mariooryad, S., Busso, C. (2013) Audiovisual corpus to analyze whisper speech. In: ICASSP. Vancouver. pp. 8101-8105.

[17] Zhang, X., Du, L., Chen, K., Zhao, X. (2007) CVSS1.0: A New Audio Visual Database For Chinese Visual Speech Synthesis. Microcomputer Applications. 28, 3: 260-265.

[18] Amini, J., Shahrebabaki, A.S., Shokouhi, N., Sheikhzadeh, H., Raahemifa, K., Eslami, M. (2013) Speech Analysis/Synthesis by Gaussian Mixture Approximation of the Speech Spectrum for Voice Conversion. In: I SSPIT. Athens. pp. 428-433.

[19] Bhuyan, A.K., Jagannath, H.N. (2015) Comparative study of voice conversion framework with line spectral frequency and mel-frequency cepstral coefficients as features using artificial neural networks. In: ICCCS. Kanyakumari. pp. 230-235.

[20] Desai, S., Black, A.W., Yegnanarayana, B., Prahallad, K. (2010) Spectral mapping using artificial neural networks for voice conversion. IEEE Transactions on, Audio, Speech, and Language Processing. 18, 5: 954-96.

[21] Chen, W., Sun, X. (2015) Mandarin Speech Emotion Recognition Based on MFCCG-PCA. Acta Scientiarum Naturalium Universitatis Pekinensis. 51, 2: 269-274.

[22] Yang C., Brown, G., Lu, L., Yamagishi, J., King, S. (2012) Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation. In: ISCSLP. Kowloon. pp. 220-223.