A Whispered Mandarin Corpus for Speech Technology Applications

Pei Xuan Lee [1], *Darren Wee*[2], *Hilary Si Yin Toh*[1], *Boon Pang Lim*[3], *Nancy Chen*[3], *Bin Ma*[3]

[1]Victoria Junior College, Singapore
[2]Anderson Junior College, Singapore
[3]Institute for Infocomm Research, Singapore

lee.pei.xuan.2011@vjc.sg, fongomong@gmail.com, toh.si.yin.hilary.2011@vjc.sg,
{bplim, nfychen, mabin}@i2r.a-star.edu.sg

Abstract

Whispered speech is a natural mode of speech in which voicing is absent – its acoustics differ significantly from normally spoken speech or so-called neutral speech, such that it is challenging to use only neutral speech to build speech processing and automatic recognition systems that can deal effectively with whisper. At the same time, humans can naturally produce and perceive whispered speech without explicit training. Tonal languages such as Mandarin present an interesting dilemma – tone is primarily encoded by pitch tracks which are absent during whispered speech, but humans can still tell tones apart. How humans manage to process whispered speech well without explicit training on it, whereas machine algorithms fail, is presently an unresolved question which could prove fruitful with study. This, however, is hindered by the lack of suitable, systematically collected corpora. We present *iWhisper-Mandarin*, a 25-hour parallel corpus of neutral and whispered Mandarin, designed to support research in linguistics and speech technology. We demonstrate and verify that earlier techniques applied to whispered speech from non-tonal languages also work with Mandarin, and present some preliminary studies on voice activity detection and whispered Mandarin speech recognition.

Index Terms: Mandarin speech processing, corpus resources, whispered speech, automatic speech recognition, voice activity detection, deep neural networks, phonation.

1. Introduction

Whispered speech is a common mode of speaking which differs from normally spoken, phonated speech (neutral speech) in that the vocal folds do not vibrate. Its acoustic-phonetic properties differ significantly from neutral speech in at least three ways[1, 2] – first, the lack of voicing and its acoustic correlates such as pitch; second, the power spectra is not only significantly weaker, the spectral tilt for whispered speech is also much gentler compared with neutral speech; third, the center frequencies for the formants also shifted from neutral speech.

Despite these differences, it is usually straightforward for a person to whisper in any language. Perception of whispered speech is also similarly straightforward in that it does not appear that there is a need to explicitly train a person to understand whispered speech, it can be done naturally. Furthermore, while to our knowledge there has not been any large scale perceptual study, anecdotally whispering does not appear to cause any significant loss in speech intelligibility. For tonal languages such as Mandarin, the lack of pitch information means the primary correlate of tone is absent in the waveform. However, past perceptual experiments have suggested that tone can still be perceived at above chance [3] [4].

All this points to a number of fundamental, unanswered questions about whispered speech which could yield fruitful insights into the nature of human speech production and perception. However, the lack of large, systematically collected whispered speech corpora continues to hinder research in this area. Recently, there has been increased interest in whispered speech, and as a result there have been a number of new speech corpora that have been collected to support interesting researches in whispered speech. Most of these, however, tend to be small or have few speakers. To our knowledge, there also does not exist a systematically collected whispered corpus of many languages including tonal ones such as Mandarin.

In this paper, we present a new Mandarin corpus developed at Institute for Infocomm Research (I^2R), as part of a larger multilingual whispered speech corpus called *iWhisper* (I^2R whispered corpus) that we plan to collect. The Mandarin section of this corpus, which we dubb *iWhisper-Mandarin* corpus, augments the earlier whispered TIMIT corpus (English) designed and collected in [5]. To our best knowledge, this is the first large-scale systematic collection of whispered speech for any tonal language. In this paper, we review existing corpora of whispered speech, describe the corpus design and collection process of *iWhisper-Mandarin*, and conduct experiments in voice activity detection (VAD) and automatic speech recognition (ASR) on iWhisper to establish baselines for future research endeavors using this new data resource. In particular, we verify that some properties that we notice about whisper speech and neutral speech observed in speech recognition experiments with Japanese and English also carry over to Mandarin.

2. Existing Corpora of Whispered Speech

A summary of whispered speech corpora is in Table 1, along with key statistics. Only whispered speech utterances are counted, we discounted any neutral speech or other modes of speech recorded in a parallel fashion. If we could not get the statistics, an estimate for the approximate number of hours of speech was made by assuming 5 seconds per utterance; these are marked with an asterisk. The column marked `P` indicates if the collected corpus is a parallel corpus where the identical sentence is both whispered and spoken neutrally.

While Table 1 is by no means comprehensive, it can be argued that even though there are corpora of whispered speech, many of these are too small to build speech recognition systems. In addition, the majority of these corpora are in English, and there is no substantial whispered corpus of Mandarin to date. We attempt to fill in this gap by presenting *iWhisper-Mandarin*. Below we give an overview of existing whispered speech corpora with an emphasis on the larger corpora that are more suitable to carry out speech recognition experiments.

2.1. Non-English Whispered Speech Corpora

One of the earliest investigations into whispered Mandarin was done by [4] in 1999. This work focused on discerning which acoustic correlates are more significant to Mandarin tone in whisper. They produced audio-visual recordings of the layrnx during whisper, and also recorded 144 isolated whispered Mandarin syllables for a perceptual test. Their analysis provides support for the theory that special maneuvers are used when whispering to preserve its intelligibility [6], namely that vocalic duration was significantly lengthened, and amplitude contour is more pronounced. However, due to the small size of the data and few number of listeners in the perceptual experiment, these findings have to be verified on a larger scale.

In 2003, Itoh and Takeda et al. [7, 2] collected Japanese neutral and whispered speech through two recording channels, and under different voicing modes and volume. A total of 68 male and 55 female speakers participated in the data collection. Each speaker reads and whispers 60 sentences of the ATR phonetically balanced sentences, and 50 sentences from the ASJ database of Japanese newspaper article sentences (JNAS). These recordings were made for both close talking microphone and mobile telephone channels, with three distinct manners of mobile handset usage. Itoh's speech recognition experiments demonstrated a peculiar asymmetry that is sometimes observed with whispered speech - with mismatched test-training conditions, a model trained using whispered speech will do better at decoding neutral speech than the converse [2].

The Whi-Spe corpus [8] developed a corpus of neutral and whispered speech in order to support their experiments on using neural networks to recognize Serbian isolated word strings. Their preliminary work using ANNs demonstrate the feasibility of performing isolated-word recognition on whispered Serbian.

2.2. English Whispered Speech Corpora

The UT-Vocal Effort (UT-VE) I and II corpora [9] is at present one of the more comprehensive collections designed to study whispered speech. UT-VE I comprises of 12 male speakers for English, recordings were made for five distinct speaking conditions based on the amount of "vocal effort" – whispered, soft, neutral, loud and shouted. Each speaker whispered at least 5 sentences selected from TIMIT sentences, as well as a minute duration of spontaneously whispered speech. The UT-VE II corpus on the other hand is a much larger corpus with large amounts of neutral speech embedded with whispered speech islands. However, while this is suitable for designing and testing systems to do voice-activity or vocal-effort detection, it is less suitable for a systematic study on the acoustic-phonetic properties of whispered speech.

In [10], the UT-VE II corpus was used to test ASR model adaptation strategies from neutral to whispered speech. They propose a modification that allows an acoustic model trained with neutral speech to decode whispered speech, doing better without any whispered data, compared to an adapted acoustic model. It would be interesting to see how well a speech recognizer trained only from whispered data performs against a neutral model adapted with limited whispered data.

The Chains corpus [11] is a speech corpus collected for the purpose of studying the effect of different speaking modes on speaker identification. It has 36 speakers recorded under various speaking conditions, including whispered and neutral speech.

The Audio-visual Whisper corpus [12], at the time of its publication is a work in progress – according to the documentation slightly more than a quarter of it has been completed. It has high-fidelity recordings of parallel whispered and neutral speech designed to study and quantify the differences during production of whisper and neutral speech

	Spkrs	Size (hrs)	P?	PB	Language
Whi-Spe [8]	5M 5F	< 5 (*)	Y	N.A.	Serbian
CIAIR [2, 7]	68M 55F	$\approx$15(*)	Y	Y	Japanese
AV-Whisper [12]	8M 3F	< 10 (*)	Y	N.A.	English
CHAINs [11]	18M 18F	< 3	Y	N.A.	English
UTVE-I [9]	12M	< 1(*)	Y	N.A.	English
UTVE-II [10]	37M 35F	< 1(*)	N	N.A.	English
wTIMIT [5]	25M 23F	≈ 15	Y	Y	English
iWhisper-Mandarin	**40M 40F**	$\approx$ **15**	**Y**	**Y**	**Mandarin**

Table 1: Whispered Speech in Research Literature Corpora. *P* stands for *parallel corpus*. *PB* stands for *phonetically balanced.*

The whispered TIMIT corpus [5] is a parallel corpus of whispered and neutral speech specifically collected and designed to complement TIMIT. It was used with various speech recognition and model adaptation schemes. In this paper we extend previous results by bringing in new, state-of-the-art DNN acoustic modeling techniques.

3. iWhisper-Mandarin: Proposed Parallel Whispered Mandarin Corpus

3.1. Corpus Description

3.1.1. Design of Reading Material

The *iWhisper-Mandarin* corpus, similar to the whispered TIMIT corpus (wTIMIT), is a parallel corpus of read and whispered speech, specifically constructed for research on whispered speech processing. We first crawled a few selected websites containing short passages and prose in Chinese, and then cleaned and parsed the data into raw text sentences. Each sentence was analyzed using an existing Mandarin pronunciation lexicon to obtain a corresponding phonetic representation. This used a greedy algorithm to segment utterances according to the longest encountered lexicon entry found in our phonetic lexicon of Mandarin words. As Mandarin is a character-based language, and the lexicon has complete entries for for up to 6500 single-characters in common use, this essentially means that there are no out-of-vocabulary words. We filtered this data to select a pool of sentences with an average of 15 characters in length. This pool was then used to generate phonetically balanced lists of 100 sentences each for speakers to read.

Each set of 100 sentences were generated in sequence, an outline of the selection algorithm for picking N speaker lists of K sentences long from a set of S sentences is given in Algorithm 1. Here, K' is a heuristic for number of candidate sentences to consider – it is used to speed up the selection algorithm. The selection algorithm works by producing a set of sentences for each speaker, one speaker at a time. Sentences are selected in a greedy fashion. Thus it manages to compile sentence lists with a phonetic distribution similar to the overall corpus, and at the same time avoids minimal repetition of sentences in the event that we do not have enough candidate sentences to build the corpus.

3.1.2. Speech Data Collection

These sentence lists were then individually read and whispered by 80 Singaporeans aged 16-20, 40 of each gender. All speakers reported an absence of articulatory and auditory impairments, and are native Mandarin speakers with over 10 years of resi-

Algorithm 1 Phonetically Balanced Corpus Construction.

function BUILDSPEAKERLISTS(S,N,K)
  $v_s \leftarrow$ COMPUTE_PHONE_VECTOR(s), $\forall s \in S$,
  $V_S = \sum_{\forall s \in S} v_s$ ▷ global vector
  $V_{tgt} = \frac{KV_S}{|S|}$ ▷ target per set
  $P \leftarrow S$ ▷ Global sentence pool
  for $u \leftarrow 1$ to N **do**
    $L_u \leftarrow 0$ ▷ sentence list
    $V_u \leftarrow 0$ ▷ set vector
    while $|L_u| < K$ **do**
      $S' \leftarrow$ RANDOM_SUBSET_OF (P, K')
      $s^* = \arg\min_{s \in S'} |V_{tgt} - (V_u + v_s)|_2$
      $L_u \leftarrow L_u \cup s^*$ ▷ add sentence
      $V_u \leftarrow V_u + v_{s^*}$ ▷ update set vector
      $P \leftarrow P - s^*$ ▷ avoid reusing sentence
      if $|P| < N'$ **then**
        $P \leftarrow P \cup S$
      end if
    end while
  end for
end function

dence in Singapore. The recordings were made with an Audio-Technica ATH-750COM USB headset microphone, sampled at 16kHz, in a quiet room with high signal-to-noise ratio. Each recorded utterance lasts between 3 to 5 seconds.

3.2. Acoustic Analysis on Vowel Length

We provide some initial acoustic analyses of the *iWhisper-Mandarin* corpus. The duration of vowel sounds have been shown to be longer in whisper [5]. We force-aligned the entire corpus to produce phone-level transcriptions, then used the minimum edit distance algorithm to align each token in a phonated utterance, with the corresponding whispered speech utterance from the same speaker. Spurious insertions or deletions due to speaker errors were discarded and only utterance pairs with identical phone tokens for both whispered and normal conditions were used in this analysis. We introduce a per utterance adjustment factor to ameliorate the effect of different speaking rate, whereby for the ith pair of aligned phones $\phi_{w,i}$ (whispered), $\phi_{n,i}$ (phonated),

$$\hat{d}(\phi_{w,i}) = d(\phi_{w,i}) \frac{\sum_{\forall i} d(\phi_{n,i})}{\sum_{\forall i} d(\phi_{w,i})},$$

$$\hat{\Delta}(\phi_i) = \hat{d}(\phi_{w,i}) - d(\phi_{n,i}). \quad (1)$$

Here $d(.)$ refers to phone duration obtained via forced-alignment, and $\hat{\Delta}$ is the adjusted change in the phone duration going from phonated to whispered speech. A positive difference suggests that the particular sound is stretched out in duration when whispered. These numbers, broken down by monophthong phone token for the major vowels is shown in Table 2.

4. Applications of the iWhisper Corpus

4.1. Voice Activity Detection on Whispered Speech

Voice activity detection (VAD) is a vital component in many speech preprocessing systems. Figure 1 illustrates a general structure used by VAD processing algorithms [13]. Most practical VAD algorithms use decision smoothing on top of a frame-level classifier – these two subsystems can be essentially decoupled for analysis. Thus, although it is crucial to properly tune

Figure 1: General Structure for VAD algorithms.

parameters for decision smoothing, in this work we focus on the accuracy of the frame-level decision module.

The decision module for recent VAD algorithms can further be categorized into two main groups – heuristic based methods relying on a log-likelihood ratio test [13], and machine learning based methods which train models to discriminate between speech and non-speech frames [14]. Heuristic approaches have traditionally been very popular and well-studied due to their simpler implementation and low overhead. Algorithms used in industry standards such as ETSI's Adaptive Multirate (AMR) and ITU's G729.B employ such heuristic VADs, which fuse different types of speech features, including full-band and sub-band energy, zero crossing rate or pitch autocorrelation.

Machine learning techniques train models that can discriminate between speech and non-speech. In [14], a support vector machine was trained with MFCC features and used as the underlying classifier – this was shown to be robust to noise. Others have also succeeded in using neural networks [15] to perform VAD. Deeper structures have also been show to provide more resilient VAD [16].

Here we only focus on machine learning approaches in this work. In particular, we replicate some experiments for VAD using neural networks on whispered speech and quantify how well a model-based approach works when training and testing on different speech modalities.

4.1.1. Experimental Setup

We used model-based speech/non-speech classifiers trained using Multilayer Perceptrons implemented by Fast Artificial Neural Network (FANN) Toolkit [17]. Our reference alignments were obtained as a by-product of the best trained speech recognition models, as described in Section 4.2. In these experiments, we used the part of the wTIMIT together with the *iWhisper-Mandarin* corpus. We selected only the portion of the speech corresponding to speakers of Singaporean English, so as to control for potential demographic biases. The wTIMIT corpus suffers from having few speakers compared to many more common speech recognition corpora. This is more pronounced if we experiment with only speakers from one demographic group. In order to ameliorate this, we performed a 96-fold cross-validation to improve the statistical reliability of our results. For each data set we drew one male and one female speaker to form a held out test set, and used the remaining speakers (8 male, 12 female) as training data. All possible combinations of speaker selections are covered by this cross-validation approach. We selected approximately one percent of the speech data from each partition to speed up training, but made sure that there were similar number of silence and non-silence frames from each speaker. This is justified by our earlier pilot experiment, as providing too much data for training will give such outstanding performance that accuracy comparisons become meaningless. This placed the number of speech frames at roughly 50,000 frames per cross-validation set.

Our setup gives us four types of source speech for training a speech-silence classifier – using neutral English, whispered English, neutral Mandarin or whispered Mandarin, and similarly four types of target speech for testing. The frontend concatenated 13-dimension MFCCs, in conjunction with their first, second, and third-order delta coefficients, 7 Funda-

PHONE	AA	OO	EE	I	AH	OH	ER	UH
Duration Increase $\hat{\Delta}$	10.3	5.2	1.4	-12	6.6	12.7	-35.9	-15

Table 2: Statistics for the average increase in phoneme duration(ms) for whispered speech.

Modality of Training speech	**Modality of Test Speech**	
	Neutral	Whispered
Neutral	96.3	83.6
Whispered	91.0	94.4

(a) Frame Classification Accuracy (English)

Modality of Training speech	**Modality of Test Speech**	
	Neutral	Whispered
Neutral	89.0	79.0
Whisper	77.2	89.8

(b) Frame Classification Accuracy (Mandarin)

Table 3: Performance of Cross-modal VAD

	Test	
Train	Eng	Man
Eng	96.3	87.5
Man	87.2	89.0

(a) Phonated Speech

Train	Test	
	Eng	Man
Eng	94.4	81.9
Man	87.8	89.7

(b) Whispered Speech

Table 4: Cross-lingual VAD performance on wTimit (English) and iWhisper (Mandarin).

mental Frequency Variation (FFV) features, and 2 Subband-Autocorrelation Classification (SAcC) features, to obtain 61-dimension feature vectors, which are fed to a single hidden-layer MLP that does speech/silence classification. For any given mode of speech and speakers for the test/train partition, the speech frames once selected, are consistently used in all remaining experiments.

4.1.2. Cross-Modal Experiments on VAD

Frame classification accuracies, averaged over all relevant speaker partitions are shown in Table 3. Here, we notice that the performance of models trained on the same speech mode is generally better, and there is some significant loss of accuracy if there is a train-test mismatch in the speech mode. For English we note that whispered speech models perform better than neutral speech model in the cross-mode testing condition. This observation is in agreement with the observation in [18] where a speech recognizer trained on whispered speech and tested on neutral speech performs significantly better than the converse system. However, this effect is not observed for the whispered Mandarin corpus.

4.1.3. Cross-Lingual Experiments on VAD

We used the trained classifiers from the first experiment, and used them to discriminate speech and silence frames of another language. A summary of the results is shown in Table 4. The numbers suggest that speech features from phonated and whispered speech differ significantly from each other, such that for even a simple classification task like speech-silence frame classification, we see a significant degradation in performance.

4.2. Speech Recognition for Whisper and Neutral speech

In this section, we present our investigations on the effectiveness of using standard training algorithms for speech recognition on whispered speech. We first trained various large vocabulary continuous speech recognizers from the data – one model for each language and each speech modality was separately trained. The training procedure for the acoustic model followed a standard recipe from the Kaldi open source toolkit [19], starting with monophone Gaussian Mixture models, and stepping up to triphones trained using Maximum-Likelihood estimation followed by Linear Discriminant Analysis. We then applied discriminative training using the Maximum Mutual Information (MMI) criterion and trained a 5 hidden layer DNN using sequential Minimum Bayes Risk (sMBR) criterion. At each stage the training data was realigned with the best trained model so far and used to train the next stage.

	Matched Test-Train			**Mismatched**	
	N+W	W	N	W → N	N→ W
MLE	23.6	29.77	15.92	46.91	91.30
DNN	12.8	17.85	10.20	34.81	64.11

Table 5: Speech recognition performance (word error rate) on neutral and whispered English speech (wTIMIT)

	Matched Test-Train			**Mismatched**	
	N+W	W	N	W → N	N→ W
MLE	61.73	66.76	54.33	81.10	98.72
DNN	53.55	60.18	50.20	71.96	93.72

Table 6: Speech recognition performance (word error rate) on neutral and whispered Mandarin Speech (iWhisper-Mandarin)

Performance over a held out test set is shown for both whispered and neutral speech in Table 5 and Table 6 for wTIMIT (English) and iWhisper-Mandarin. Our contribution to the literature is to demonstrate that by using state-of-the-art acoustic modeling techniques such as using Deep Neural Networks trained with sequential Minimum Bayes Criterion (DNN-sMBR), a substantial increase in recognition performance can be attained. The model appears to be able to capture enough saliency about speech to be reasonably robust to mismatched speaking modes. In contrast, discriminative training using MMI can hurt the performance in mismatched recognition. Overall the numbers reaffirm the observations in [2] and [5], indicating that a speech recognizer trained from whispered speech can still recognize neutral speech reasonably, but not vice versa. In addition, we observe that the performance of the Mandarin system is consistently worse than that of the corresponding English system for all conditions. We plan to further investigate this discrepancy with future work.

5. Discussion

We have presented the *iWhisper-Mandarin* corpus, a 25-hour parallel corpus of phonated and whispered speech collected over 80 Mandarin speakers. We demonstrate that classical VAD and speech recognition algorithms can perform reasonably well even on whispered speech alone. We plan to further study the performance of automatic tone recognition in whispered Mandarin. We also intend to make *iWhisper-Mandarin* publicly available in the near future.

6. References

[1] S. T. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," pp. 263–274, 2008.

[2] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, pp. 139–152, October 2003.

[3] A. S. Abramson, "Tonal experiments with whispered Thai," in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, A. Valdman, Ed. The Hague, Netherlands: Mouton, 1972, pp. 31–44.

[4] M. Gao, "Tones in whispered chinese: Articulatory features and perceptual cues," Master's thesis, 1999.

[5] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana Champaign, Dec 2010.

[6] W. Meyer-Eppler, "Realization of prosodic features in whispered speech," *Journal of the Acoustical Society of America*, vol. 29, no. 1, 1956.

[7] N. Kawaguchi, K. Takeda, S. Matsubara, I. Yokoo, T. Ito, K. Tatara, T. Shinde, and F. Itakura, "Ciair speech corpus for real world applications," in *COCOSDA*, 2002.

[8] D. T. Grozdic, B. Markovic, J. Galic, and S. T. Jovocic, "Application of neural networks in whispered speech recognition," 2013.

[9] "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," vol. 19, no. 4, May 2011.

[10] H. B. Shabnam Ghaffarzaegan and J. H. L. Hansen, "Utvocal effort ii: Analysis and constrained lexicon recognition of whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[11] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: Characterizing individual speakers," in *SPECOM*, St Petersburg, Russia, 2006, pp. 421–435.

[12] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *ICASSP*, 2013.

[13] J. Ramirez, J. M. Gorriz, and J. Segura, "Voice activity detection: Fundamentals and speech recognition system robustness," in *Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. University of Granada, Spain: I-TECH, 2007.

[14] T. Kinnunen, E. Chernenko, M. Tuononen, P. Frunti, and H. Li, "Voice activity detection using mfcc features and support vector machine," 2007.

[15] T. V. Pham, C. T. Tang, and M. Statschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," 2009.

[16] N. Ryant, M. LIberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks," 2013.

[17] S. Nissen, "Implementation of a Fast Artificial Neural Network Library (fann)," *Report, Department of Computer Science University of Copenhagen (DIKU)*, vol. 31, 2003.

[18] T. Itoh, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *Acoustics Speech and Signal Processing*, vol. 1, 2002, pp. 389–392.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vasely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.