AISHELL6-WHISPER: A CHINESE MANDARIN AUDIO-VISUAL WHISPER SPEECH DATASET WITH SPEECH RECOGNITION BASELINES

Cancan Li[1,2], Fei Su[1,2], Juan Liu[2,1†], Hui Bu[4], Yulong Wan[5], Hongbin Suo[5], Ming Li[2,3†]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]School of Artificial Intelligence, Wuhan University, Wuhan, China
[3]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Digital Innovation Research Center,
Duke Kunshan University, Kunshan, China
[4]Beijing AISHELL Technology Co., Ltd., Beijing, China
[5]AI Center, OPPO, Beijing, China

ABSTRACT

Whisper speech recognition is crucial not only for ensuring privacy in sensitive communications but also for providing a critical communication bridge for patients under vocal restraint and enabling discrete interaction in noise-sensitive environments. The development of Chinese mandarin audio-visual whisper speech recognition is hindered by the lack of large-scale datasets. We present AISHELL6-Whisper, a large-scale open-source audio-visual whisper speech dataset, featuring 30 hours each of whisper speech and parallel normal speech, with synchronized frontal facial videos. Moreover, we propose an audio-visual speech recognition (AVSR) baseline based on the Whisper-Flamingo framework, which integrates a parallel training strategy to align embeddings across speech types, and employs a projection layer to adapt to whisper speech's spectral properties. The model achieves a Character Error Rate (CER) of 4.13% for whisper speech and 1.11% for normal speech in the test set of our dataset, and establishes new state-of-the-art results on the wTIMIT benchmark. The dataset and the AVSR baseline codes are open-sourced at https://zutm.github.io/AISHELL6-Whisper.

Index Terms— Whisper speech recognition, Audio-visual speech recognition, Audio-visual whisper speech dataset

1. INTRODUCTION

Whisper speech typically occurs in private conversations under quite and public acoustic environments or in pathological conditions, such as vocal cord damage [1–3]. Automatic whisper speech recognition is valuable not only for enabling discreet communication in public settings but also for its potential medical applications. Unlike normal speech, whisper speech is produced without vocal cord vibration [4], resulting in the absence of fundamental frequency and generally lower energy [5, 6]. These characteristics make the accurate recognition of whisper speech particularly challenging [7].

When collecting the whisper speech using the top microphone of the cell phone, the front-facing camera can also simultaneously capture the lip movement which can be quite useful for speech recognition. However, audio-visual whisper speech recognition remains relatively underexplored, largely due to the scarcity of dedicated datasets. For Chinese Mandarin, only a few small-scale whisper speech corpora are publicly available. The iWhisper-Mandarin dataset [8] was the first released resource, containing approximately

25 hours of whisper and normal recordings for automatic speech recognition (ASR). In [9], the first Chinese audio-visual whisper dataset comprising approximately 2.44 hours of recordings from 10 speakers is introduced.

To advance whisper speech recognition, early models [5, 7] utilized pseudo-whisper data based on Gaussian Mixture Model-Hidden Markov Models (GMM-HMM) to improve performance, while another study [10] employed pseudo-whisper data based on Deep Neural Networks-Hidden Markov Models (DNN-HMM). Chang et al. [3] proposed an end-to-end (E2E) whisper recognition system, incorporating a frequency-weighted SpecAugment method and a frequency-divided Convolutional Neural Networks (CNN) extractor. Subsequently, Lin et al. [2] enhanced E2E performance through pseudo-whisper data augmentation, while Farhadipour et al. [1] fine-tuned a pre-trained WavLM model [11] for both whisper and dialect whisper speech recognition. These studies have demonstrated the potential of leveraging deep learning to improve whisper recognition. However, the model structures used in these works still lag behind the performance of advanced speech recognition models published in recent years. Additionally, the effectiveness of training with pseudo-whisper data requires further verification, as it differs from real whisper data and may confuse the model.

In this work, we introduce the AISHELL6-Whisper dataset to address this gap. It is the largest publicly available audio-visual whisper speech dataset, containing approximately 30 hours each of high-quality whisper speech and parallel normal speech, with synchronized frontal facial videos.

We also implemented Whisper-Flamingo [12] as the baseline of our audio-visual whisper speech recognition systems. This model incorporates visual features from AV-HuBERT [13] into the OpenAI Whisper speech recognition and translation framework [14] through gated cross-attention [15], achieving state-of-the-art performance in audio-visual speech recognition on the LRS3 [16] dataset. We employ a parallel training strategy to align whisper speech embeddings with normal speech embeddings together and apply an additional projection layer on top of the whisper speech features to further improve the recognition accuracy of whisper speech.

2. DATASET

2.1. Data Collection

The AISHELL6-Whisper corpus was collected in a controlled studio environment, containing parallel recordings of whisper and normal

---

Table 1. Statistics for the Train, Valid and Test sets of AISHELL6-Whisper. 'N' indicates normal speech and 'W' indicates whisper speech.

Set	Video	Num of Spk	Type	Time (hrs)	Utterances
Train	Yes	82	N	14.8339	10012
			W	14.8359	9971
	No	28	N	4.8887	3335
			W	5.0526	3332
Valid	Yes	19	N	3.3252	2331
			W	3.3983	2304
	No	10	N	1.7310	1190
			W	1.8038	1191
Test	Yes	20	N	3.4126	2422
			W	3.3601	2351
	No	8	N	1.3477	938
			W	1.3021	938

Table 2. Comparison with existing whisper speech datasets. "Parallel" indicates the availability of both whisper and normal speech recordings, "Time (hrs)" refers to the duration of whisper speech.

Dataset	Type	Language	Time(hrs)	Parallel	Num of Spk
UTVE-I [17]	ASR	English	<1	✓	12
UTVE-II [18]	ASR	English	1	-	112
AVWD [9]	AVSR	Chinese	<2.44	✓	10
CHAINs [19]	ASR	English	<3	✓	36
Whi-Spe [20]	ASR	Serbian	<5	✓	10
AV-Whisper [21]	AVSR	English	<10	✓	11
CIAIR [22]	AVSR	Japanese	15	✓	123
iWhisper-Mandarin [8]	ASR	Chinese	15	✓	80
wSPIRE [23]	ASR	English	18	✓	88
wTIMIT [24]	ASR	English	26	✓	48
AISHELL6-Whisper	AVSR	Chinese	30	✓	167

speech. It comprises 167 speakers, each reading approximately 10-20 minutes of poetry texts without any overlap in content. Ground-truth transcriptions were manually annotated. Among them, 121 participants were recorded using both a high-fidelity microphone and a synchronized RGBD camera, while the remaining 46 participants only recorded the audio signals. Audio was captured with a single-channel high-fidelity microphone (Neumann U87) at a 48 kHz sampling rate, with a background noise level of less than 20 dB. The microphone was positioned below the speaker's chin to ensure sound quality without obscuring the speaker's lip movement. Video recordings were captured using a RGBD camera placed one meter directly in front of the speaker, with a resolution of 1280×720 at 25 fps.

We divided the dataset into training, validation, and test subsets using an approximate 4:1:1 ratio, ensuring a balanced distribution of age and gender across splits. There are no speaker overlaps between these three subsets. The dataset statistics are summarized in Table 1.

A comparison with other existing whisper speech datasets is provided in Table 2. Our dataset features relatively larger size, with parallel recording setup and unique multimodal design.

2.2. Data Processing

To extract lip movements from the dataset, facial landmarks were extracted using RetinaFace [25]. Subsequently, lip regions of each frame are cropped following the methodology in [26] using facial
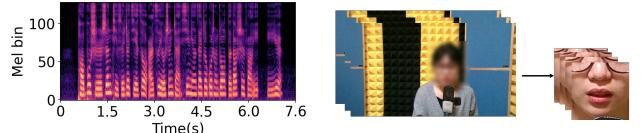

(a) Mel spectrogram of a normal speech.

(b) Lip movement processing of the normal speech
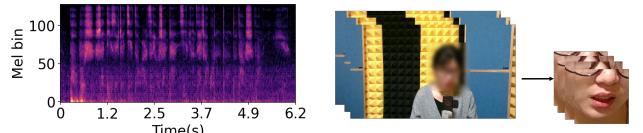
(c) Mel spectrogram of the corresponding whisper speech.

(d) Lip movement processing of the whisper speech

Fig. 1. Data view of AISHELL6-Whisper with speaker ID "0120" and utterance "跑步时身体随着节奏而自由伸展心灵在这过程中得到释放与愉悦".

landmarks. The mouth center is defined as:

$$x_{center}, y_{center} = \frac{p_2 + p_3}{2} \tag{1}$$

where $p2$ and $p3$ represent the left mouth corner and the right mouth corner, respectively. The width of the cropping bounding box is determined by:

$$width = min\{3.2 \times d_{MN}, 2 \times max\{d_{MN}, d_{p1p2}\}\} \tag{2}$$

where d_{p1p2} denotes the distance between $p2$ and the nose point $p1$, and d_{MN} is the Euclidean distance between $p1$ and the mouth center. Lip movement videos were generated by extracting (96,96) lip regions.

Since speakers may occasionally make minor errors while reading texts, whisper speech may exhibit slight differences from normal speech when reading the same sentence. To address these variations, we calculate the textual similarity for each whisper segment against all normal speech segments from the same speaker using the SequenceMatcher[1] algorithm. The normal speech segment from the same speaker and with the highest similarity score is selected as the matching pair.

Fig.1 provides an example of the processed audio and video data, illustrating mel spectrograms for normal speech and the corresponding whisper speech, along with visualized lip movements corresponding to the audio. The significant spectral differences observed in the speech spectrogram-particularly the absence of fundamental frequency-demonstrate the acoustic challenges that traditional speech recognition systems face when processing whisper speech.

3. THE PROPOSED BASELINE MODEL

3.1. Model structure

We train the model in two steps, following the methodology introduced in Whisper-Flamingo [12].

In the first step, the OpenAI Whisper model [14] is finetuned on the audio data of whisper and normal speech. Our approach employs a parallel training strategy that processes paired utterances of

---

[1] https://docs.python.org/3/library/difflib.html
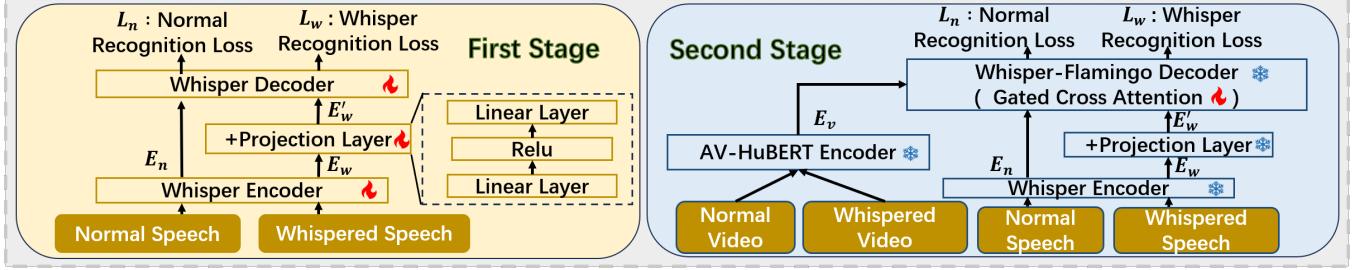
Fig. 2. The architecture of our proposed baseline model. In the first stage, paired utterances of whisper and normal speech are simultaneously processed through a shared Whisper encoder, with a projection layer module refining the whisper embedding. Both embeddings are fed into the decoder to calculate the cross-entropy loss between the model's predicted transcripts and the ground-truth tokens independently. The second stage follows the Whisper-Flamingo [12] approach, incorporating visual features via gated cross-attention for audio-visual training.

Table 3. Speech recognition performance on the AISHELL6-Whisper testing data. AISHELL6-Whisper training set is used for finetuning and training models.

Model	Parallel training	Projection layer	Video	Whisper Speech CER	Normal Speech CER
Whisper (Large-V3)	-	-	-	18.93%	3.95%
+Finetune	-	-	-	6.69%	1.62%
+Parallel training	✓	-	-	4.53%	**0.98%**
+Projection layer	✓	✓	-	4.34%	1.14%
+Video	✓	-	✓	4.21%	1.08%
+Video (Proposed)	✓	✓	✓	**4.13%**	1.11%

whisper and normal speech simultaneously, leveraging their intrinsic relationship to enhance model's performance on both speech types, as described in the first stage of Fig.2. The Whisper encoder and decoder process paired utterances simultaneously, optimizing the recognition losses for both whisper and normal speech with the following objective:

$$\mathcal{L} = \mathcal{L}_w + \mathcal{L}_n \tag{3}$$

where $\mathcal{L}_w$ and $\mathcal{L}_n$ represent the standard cross-entropy losses between the model's predicted transcripts and ground-truth tokens for whisper and normal speech, respectively.

In the second step, the model is further finetuned using audio-visual data, following the structure of the second stage in Fig.2. In this stage, a gated cross-attention module is integrated into the beginning of each decoder block in Whisper, resulting in the Whisper-Flamingo Decoder. At this training stage, only the gated cross-attention module is trained, allowing the model to incorporate visual input effectively. Both videos corresponding to the normal speech and whisper speech are fed into the AV-HuBERT Encoder. For data without video signals, video inputs are set to zero.

3.2. Projection Layer

To better capture the information absent in whisper speech compared to normal ones, we add a lightweight projection layer module with a linear structure (Linear → Rectified Linear Unit (ReLU) → Linear) after the Whisper encoder. The two linear components in the structure maintain identical input and output dimensions. It serves to refine the whisper speech embeddings, aligning them more closely with the acoustic characteristics of normal speech, thereby effectively bridging the gap for whisper speech. The refinement process is formulated as follows:

$$E'_w = E_w + \text{projection layer}(E_w) \tag{4}$$

where E_w represents the original whisper speech embedding from the encoder and E'_w is the refined embedding.

The projection layer is carefully initialized to ensure stable training. The first linear layer employs Kaiming normal initialization [27], suitable for the subsequent ReLU activation function. The final linear layer is initialized to zeros, resulting in an identity transformation at the start of training. This initialization strategy allows the model to start from a stable baseline and gradually learn the necessary transformations.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Database and Evaluation Metrics

To measure the model's performance on existing benchmarks, it is evaluated on the wTIMIT corpus [24], which contains 450 phonetically balanced sentences spoken in both normal (wTIMIT-n) and whisper (wTIMIT-w) speech by 48 speakers with US and Singaporean (SG) English accents. Each speech type, either whisper or normal, provides approximately 26 hours of data. The current train-test split of wTIMIT contains a significant overlap in utterances, with the same sentences spoken by different speakers, making it unsuitable for accurately estimating recognition accuracy [3]. Partitioning by speakers only reduced performance by approximately 10% compared to the case with speaker overlap [3], and partitioning by speakers can result in less data available for training [2]. Therefore, prohibiting speaker overlap between subsets is unnecessary [2, 3]. To ensure a consistent comparison with prior works, the dataset is repartitioned in accordance with [2, 3]. The resulting splits consist of 400 sentences for training and 25 sentences each for validation

Table 4. Comparison of WER and CER (%) in the wTIMIT testing data. "Pre-trained" indicates models fine-tuned using checkpoints pre-trained on AISHELL6-Whisper training set data. "Proposed Strategies" indicates using parallel training and projection layer. "N" indicates normal speech and "W" indicates whisper speech, with "US" stands for US accents and "SG" stands for Singapore accents. Chang et al. [3] evaluated CER on the entire whisper test set and the normal test set, so we provide results that lie between those of the US and SG sets.

Model	Pre-trained	Proposed Strategies	WER (%)				CER (%)			
			N_{US}	N_{SG}	W_{US}	W_{SG}	N_{US}	N_{SG}	W_{US}	W_{SG}
[3]	-	-	—	—	—	—	23.5		19.0	
[2]	-	-	26.8	38.6	30.7	49.2	—	—	—	—
Whisper (Large-V3)	-	-	5.25	8.67	8.20	19.03	1.03	2.50	2.47	8.85
+ Finetune	-	-	5.58	7.46	7.95	15.79	0.93	1.86	2.37	7.53
+ Finetune*	✓	-	5.78	7.64	8.42	13.84	0.95	1.81	2.15	5.76
+ Finetune (Proposed)*	✓	✓	**4.50**	**6.06**	**6.35**	**11.63**	**0.73**	**1.33**	**1.85**	**5.17**

and testing for each speaker.

For evaluation, Word Error Rate (WER) and Character Error Rate (CER) are used for the English wTIMIT corpus, while CER is employed for our Chinese speech recognition task.

4.2. Experimental Setup

The language identification parameter is set to "zh" and "en" for Chinese and English speech recognition tasks, respectively.

All experiments are based on the pre-trained OpenAI Whisper Large-v3 model [14]. A maximum audio length of 320,000 samples (20 seconds) is enforced during both training stages. The audio is resampled to 16 kHz to match the Whisper model's input specification. All other hyperparameters and configurations are adhered to those established in the original Whisper-Flamingo framework [2] for fine-tuning the large Whisper and AV-HuBERT models.

In the first stage, the Whisper model is trained for two epochs. In the second stage, the model is further trained with lip movement videos for four epochs.

4.3. Experimental results

4.3.1. Performance on AISHELL6-Whisper

Table 3 presents the CER in the AISHELL6-Whisper testing data. The original Whisper model exhibits significantly degraded performance on whisper speech, with a CER of 18.93%, which stands in stark contrast to the 3.95% CER achieved on normal speech, highlighting the inherent challenge in whisper speech recognition.

Fine-tuning Whisper on the AISHELL6-Whisper training dataset significantly reduces the CER for whisper speech to 6.69% and normal speech to 1.62%. Our proposed parallel training strategy further improves performance across both modalities, lowering the whisper CER to 4.53% and the normal CER to 0.98%, by enabling mutual reinforcement and compensating for acoustic differences.

The introduction of a lightweight projection layer, specifically designed for whisper speech, further optimizes performance for the target modality, achieving a CER of 4.34% on whisper speech. As expected, this specialization leads to a slight performance trade-off in normal speech, with the CER increasing to 1.14%.

Incorporating video information provides complementary visual cues. For the model without the projection layer, video integration enhances whisper recognition (reducing CER from 4.53% to 4.21%) but results in a minor regression in normal speech recognition (increasing CER from 0.98% to 1.08%). This suggests a shift in the model's focus toward leveraging the visual modality. Notably, when

combined with the projection layer, video integration results in optimal performance, achieving the lowest whisper CER of 4.13% while maintaining a low normal CER of 1.11%. This configuration demonstrates the efficacy of our multimodal approach, offering an optimal balance for recognizing both types of speech. Future studies can explore the usage of the video modality under noisy and low SNR scenarios in which visual cues becomes more important.

4.3.2. Performance on wTIMIT

To evaluate the generalizability of the proposed techniques, we conducted experiments on the wTIMIT dataset. The wTIMIT training set's limited size (400 sentences for each speaker) increases the risk of overfitting, which makes it difficult for the model to learn whisper-specific features effectively. To address this and improve learning, we initialized our model using checkpoints pre-trained on AISHELL6-Whisper dataset with parallel training and projection layer strategies, before fine-tuning on wTIMIT. Since wTIMIT only includes audio data, we only trained on the first stage of our proposed model configuration.

As shown in Table 4, substantial performance gains are achieved on both whisper speech and normal speech. The final model, which is fine-tuned using the proposed parallel training and projection layer strategies, demonstrates statistically superior performance to the directly fine-tuned baseline. Specifically, for US accents, the WER for whisper speech decreases by 1.85% compared to the original Whisper Large-v3 model. For Singaporean accents (SG), the WER for whisper speech is reduced by a remarkable 7.40%, demonstrating that our model not only performs well on our own AISHELL6-Whisper test data in Chinese but also exhibits exceptional transferability to other datasets in English.

5. CONCLUSION

This work contributes to the field of whisper speech recognition by introducing AISHELL6-Whisper, the largest audio-visual whisper speech dataset in Chinese Mandarin, comprising 30 hours each of whisper speech and parallel normal speech, with synchronized frontal facial videos.

The proposed audio-visual speech recognition baseline, integrating the parallel training strategy with the projection layer based on the Whisper-flamingo architecture, demonstrates superior performance with a CER of 4.13% on whisper speech and 1.11% for normal speech in our test set while achieving state-of-the-art results on the public wTIMIT benchmark.

---

[2] https://github.com/roudimit/whisper-flamingo

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Aref Farhadipour, Homa Asadi, and Volker Dellwo, "Leveraging self-supervised models for automatic whispered speech recognition," in *Proc. ICCKE*. IEEE, 2024, pp. 188–193.

[2] Zhaofeng Lin, Tanvina Patel, and Odette Scharenborg, "Improving whispered speech recognition performance using pseudo-whispered based data augmentation," in *Proc. ASRU*. IEEE, 2023, pp. 1–8.

[3] Heng-Jui Chang, Alexander H Liu, Hung-yi Lee, and Lin-shan Lee, "End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training," in *Proc. SLT*. IEEE, 2021, pp. 186–193.

[4] Slobodan T Jovičić and Zoran Šarić, "Acoustic analysis of consonants in whispered speech," *Journal of voice*, vol. 22, no. 3, pp. 263–274, 2008.

[5] Shabnam Ghaffarzadegan, Hynek Bořil, and John HL Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *Proc. TASLP*, vol. 24, no. 10, pp. 1705–1720, 2016.

[6] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura, "Analysis and recognition of whispered speech," *Speech communication*, vol. 45, no. 2, pp. 139–152, 2005.

[7] Đorđe T Grozdić and Slobodan T Jovičić, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *Proc. TASLP*, vol. 25, no. 12, pp. 2313–2322, 2017.

[8] Pei Xuan Lee, Darren Wee, Hilary Si Yin Toh, Boon Pang Lim, Nancy F Chen, and Bin Ma, "A whispered mandarin corpus for speech technology applications.," in *Proc. Interspeech*, 2014, pp. 1598–1602.

[9] Jian Zhou, Yuting Hu, Hailun Lian, Cong Pang, Huabin Wang, and Liang Tao, "An audio-visual whisper database in chinese," in *Journal of Physics: Conference Series*. IOP Publishing, 2019, vol. 1237, p. 022106.

[10] Shabnam Ghaffarzadegan, Hynek Bořil, and John HL Hansen, "Deep neural network training for whispered speech recognition using small databases and generative model sampling," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 1063–1075, 2017.

[11] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[12] Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass, "Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation," *Proc. Interspeech*, pp. 2420–2424, 2024.

[13] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *Proc. ICLR*, 2022.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*. PMLR, 2023, pp. 28492–28518.

[15] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al., "Flamingo: a visual language model for few-shot learning," *Proc. NeurIPS*, vol. 35, pp. 23716–23736, 2022.

[16] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv:1809.00496*, 2018.

[17] Chi Zhang and John HL Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 883–894, 2010.

[18] Shabnam Ghaffarzadegan, Hynek Bořil, and John HL Hansen, "Ut-vocal effort ii: Analysis and constrained-lexicon recognition of whispered speech," in *Proc. ICASSP*. IEEE, 2014, pp. 2544–2548.

[19] F Cummins, M Grimaldi, T Leonard, and J Simko, "The chains corpus: Characterizing individual speakers," 2006.

[20] Đorđe T Grozdić, Branko Marković, Jovan Galić, and Slobodan T Jovičić, "Application of neural networks in whispered speech recognition," in *Proc. TELFOR*. IEEE, 2012, pp. 728–731.

[21] Tam Tran, Soroosh Mariooryad, and Carlos Busso, "Audio-visual corpus to analyze whisper speech," in *Proc. ICASSP*. IEEE, 2013, pp. 8101–8105.

[22] 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 *Proc. COCOSDA*, 2002.

[23] Bhavuk Singhal, Abinay Reddy Naini, and Prasanta Kumar Ghosh, "wspire: A parallel multi-device corpus in neutral and whispered speech," in *Proc. O-COCOSDA*. IEEE, 2021, pp. 146–151.

[24] Boon Pang Lim, *Computational differences between whispered and non-whispered speech*, University of Illinois at Urbana-Champaign, 2011.

[25] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proc. CVPR*, 2020, pp. 5203–5212.

[26] Ming Cheng, Haoxu Wang, Yechen Wang, and Ming Li, "The dku audio-visual wake word spotting system for the 2021 misp challenge," in *Proc. ICASSP*. IEEE, 2022, pp. 9256–9260.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.