INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Effects of base-frequency and spectral envelope on deep-learning speech separation and recognition models

J. Hui[1,2], *Y. Wei*[1,3], *S.T. Chen*[1,2], *and R.H.Y. So*[1,2]

[1]HKUST-Shenzhen Research Institute, Shenzhen, China
[2]Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong, China
[3]Bioengineering Graduate Program, School of Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

jhuiac@connect.ust.hk, yweiaj@connect.ust.hk, schenbq@connect.ust.hk, rhyso@ust.hk

Abstract

Base-frequencies (F0) and spectral envelopes play an important role in speech separation and recognition by humans. Two experiments were conducted to study how trained networks for multi-speaker speech separation/recognition are affected by difference of F0 and spectral envelopes between source signals. The first experiment examined the effects of natural F0/envelope on the performance of speech separation. Results showed that when the two target signals differed in F0 by ±3 semitones or more or differed in the envelope by a scaling factor larger than 1.08 or less than 0.92, separation performance improved significantly. This is consistent with human listeners and is the first finding for deep learning-network (DNN) models. The second experiment tested the effect of F0/envelope difference on multi-speaker automatic speech recognition(ASR) system's performance. Results showed that multi-speaker recognition result also significantly rely on F0/envelope differences. The overall results indicated that the dependency of the existing automatic systems on monaural cues is similar to that of human, while automatic systems still perform inferior than human on same tasks.

Index Terms: speech separation, speech recognition, base frequency, envelope

1. Introduction

Humans have the ability to focus on a single speaker while filtering out all other sounds. This phenomenon is known as the cocktail party effect (CPE) and can be difficult for machines to imitate [1], [2], [3], [4]. The CPE has been the subject of many studies [5], [6], [7], [8], [9], [10], [11]. Researchers in the fields of psychophysics and neuroscience have discovered that besides binaural spatial cues, monaural cues also play an important role in CPE tasks [7], [12]. In particular, evidence suggests that listeners can rely solely on monaural cues to recognize a target speech in the presence of a masking speech. Past experiments showed that differences in the spectral envelope and base frequency—also referred to as the fundamental frequency (F0)—between concurrent speech streams are the two most influential monaural cues [7], [8], [9], [13]. F0 is determined by vocal cord vibration while the configuration of the vocal tract determines the spectral envelope [7], [14]. Darwin and his colleagues reported that experimental participants were able to discern what two individuals speaking simultaneously were saying when the difference in F0 between the two speakers exceeded 2 semitones [7]. With the same F0, speech recognition performance improved when the ratio of the vocal tract lengths of the two speakers was 1.08 or greater [7]. However, similar studies on automatic speech recognition (ASR) models could not be found.

Knowledge of the important role of F0 and the spectral envelope has accelerated the development of speech separation methods such as computational auditory scene analysis (CASA) [15], nonnegative matrix factorization (NMF) [16], and model-based methods [17]. In recent years, deep learning (DL)-based speech separation systems have made great progress [15], [18], [19], [20], [21], [22]. Most studies on DL-based ASR focus on optimizing certain empirical performance indexes such as the signal-to-distortion ratio (SDR) and the word error rate (WER), while ignoring the effects of basic parameters such as F0 and the spectral envelope.

This paper reports the findings of two experiments using 10 DL-based multi-speaker ASR systems as "listeners" to examine the effects of F0 and the spectral envelope. Insights on how to improve the current speech separation and multi-speaker ASR systems will be discussed.

2. Experiment 1: Speech Separation

2.1. Dataset

The publicly available speech corpus CSTR Voice Cloning Toolkit (VCTK) was used [23]. This corpus consisted of speech uttered by 109 native English speakers with various accents and has been used in other studies [24], [25]. Each speaker spoke about 400 sentences. We constructed our experimental dataset by randomly selecting five male speakers and five female speakers with different accents from the corpus.

All utterances of the selected speakers were down-sampled from 48kHz to 16kHz after passing through anti-aliasing filters for consistency with past studies (see [13], [20], [24], [25] for studies using a sampling rate of 16kHz and [18], [19], [21], [26], [27], [28] for studies using 8kHz). We also ran all analyses using a sampling rate of 20kHz and the results showed no difference. The utterances were processed with the speech synthesis package WORLD [26] to produce seven new sets of speech files with F0 contours shifted by different numbers of semitones. In order to keep the modified F0 in the typical range for human speech, the utterances by female speakers were shifted by -9, -6, -3, -1, 0, +1 and +3 semitones while those by male speakers were shifted by -3, -1, 0, +1, +3, +6 and +9 semitones. In the following paragraphs, we use ΔF0 to represent the change in F0.

Similarly, we also produced seven new sets of speech files

http://dx.doi.org/10.21437/Interspeech.2019-1715

with the spectral envelope shifted using the envelope shifting method provided in the WORLD package [26]. Specifically, an envelope was shifted by a factor α by (1) generating a smoothed spectrogram from the speech signal and eliminating the influence of F0; (2) normalizing the frequency axis of the spectrogram; (3) calculating the $1/\alpha$ power of the normalized frequency axis as the new coordinate of the normalized frequency axis; and (4) recovering the frequency axis of the spectrogram. In this way, when α exceeded 1, the low-frequency components were extended along the frequency axis and the high-frequency components were compressed. As a result, energy was more concentrated in the high-frequency region (since the energy of human speech is mainly concentrated in the region below 3000Hz [29], the changing effect on the high-frequency components can be ignored). When α was less than 1, the effect was reversed. The advantage of this envelope shifting method is that the range of the spectrogram's frequency axis remains unchanged. In order to keep modified voices in the range of normal human speech [13], the utterances of female speakers were shifted by α values of 0.84, 0.88, 0.92, 0.96, 1, 1.04 and 1.08, and those of male speakers were shifted by 0.92, 0.96, 1, 1.04, 1.08, 1.12 and 1.16.

For each ΔF0/α setting, we constructed 500 speech mixtures in the training set and another 500 in the test set. The training set and the test set did not overlap with each other. Each speech mixture contained two randomly selected sentences spoken by the same speaker. One was unchanged and the other had a shifted F0/envelope, with the signal-to-noise ratio (SNR) selected uniformly between -2.5dB and 2.5dB.

2.2. Approach

The deep learning model used was the permutation invariant training speech separation (PIT-SS) model proven to be able to imitate the CPE [20], [27], [28]. The PIT-SS model in this article has a setup similar to that in [28] but with bidirectional long short-term memory (LSTM) with 128 memory cells in each layer. The input to the model was the 129-dim short-time Fourier transform (STFT) spectral magnitude of the speech mixture, computed using the STFT with a frame size of 32ms and a 16ms shift. The output layer was divided into two output streams, each of which belonged to one real or synthesized speaker.

We trained 10 PIT-SS models (cf. 10 listeners in an experiment) for each of the 140 ΔF0-α values ($140 = 10$speakers $\times$ $(7\Delta\text{F0} + 7\alpha)$) for a total of 1,400 trials.

2.3. Results

2.3.1. Effects of shifting F0

The metric we used to evaluate speech separation results was the signal-to-distortion ratio (SDR), which has been widely used to evaluate speech enhancement performance [30].

Inspection of FIG 1 indicates that the performance of separating the male voice increased monotonically with differences in F0 while that of separating the female voice exhibited an inverted U shape. Future work is needed to determine whether the performance for male speakers would also exhibit an inverted U shape if F0 is shifted for more than -3 semitones.

In order to eliminate differences among individual speakers, and highlight differences between genders, we plotted the average performance for five female speakers and five male speakers in FIG 2. Changing ΔF0 from +1 to +3 semitones or from -1 to -3 semitones improved the SDR by 4.1 or 2.4dB for female speakers ($p < 0.001$). For male speakers, the separation performance improved by 2.1 dB when ΔF0 was increased from +1 to +3 semitones ($p < 0.001$).

As the absolute value of ΔF0 increased, the overall performance of speech separation increased asymptotically (see FIG 1).

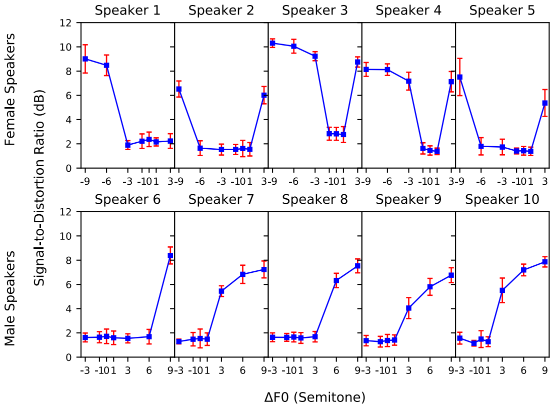

Figure 1: *Speech separation performance (SDR) as a function of* Δ*F0 in semitones. Data from each "speaker" are shown. The error bars represent the* 95% *confidence intervals. F0 was shifted in different directions to maintain the normal human tune.*

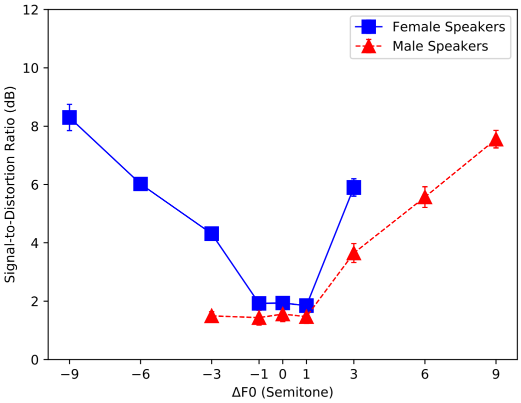

Figure 2: *Average separation performance for female speakers and male speakers as a function of* Δ*F0. The error bars represent the 95% confidence intervals.*

2.3.2. Effects of shifting the spectral envelope

The speech separation performance for each speaker is illustrated as functions of shifts in the spectral envelope in FIG 3.

FIG 4 shows the average speech separation performance as a function of α by gender. When α was reduced from 0.96 to 0.92 or increased from 1.04 to 1.08, the SDR performance was significantly improved by 3.4dB or 2.4dB respectively for female voices ($p < 0.001$) and by 4.5dB or 3.6dB for male voices ($p < 0.001$).

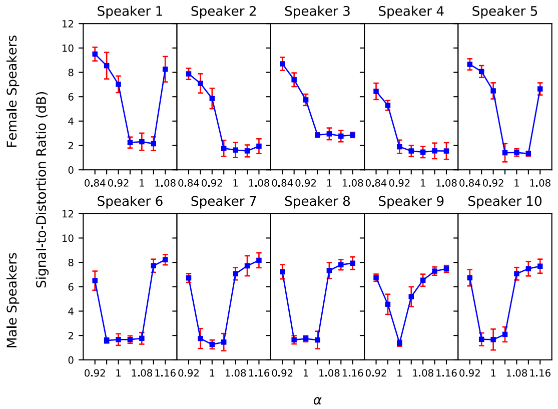

Figure 3: *SDR of separated sentences as a function of envelope scaling factor α, where each subfigure belongs to one speaker. The error bars represent the 95% confidence intervals.*

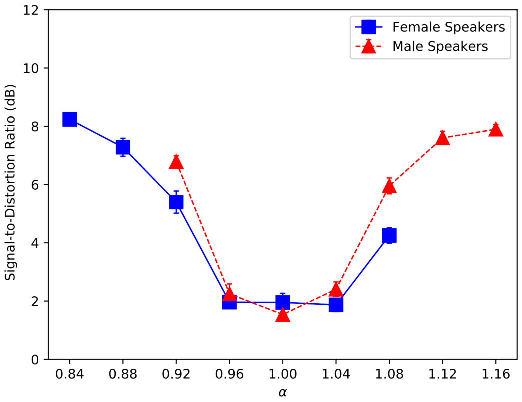

Figure 4: *Average separation performance for female speakers and male speakers as a function of α. The error bars represent the 95% confidence intervals.*

3. Experiment 2: Speech Recognition

3.1. Dataset

Similar to Experiment 1, we also used the VCTK dataset and the values of ΔF0, α and the speaker sets were the same as in Experiment 1.

3.2. Approach

This experiment proceeded in two steps: first, we tested the effect of shifting F0 and the envelope on clean speech automatic recognition; second, we tested the influence of changing F0 and the envelope on the joint multi-speaker ASR system comprising the PIT-SS system and the clean speech automatic recognition system. In this experiment, a pre-trained model, Google Speech Recognition API provided in the Speech Recognition package [31] was used as our clean speech automatic recognition model, and the best possible performance was obtained by choosing the right accent for certain speakers.

3.3. Results

3.3.1. *Effects of shifting F0 and the envelope on clean speech automatic recognition*

The word error rate (WER) is a common metric for the performance of a speech recognition system and was used to evaluate the results of clean speech automatic recognition. Effects of changes in the base frequency and spectral envelope on the performance of clean speech automatic recognition are plotted in FIG 5 and FIG 6 respectively.

Interestingly, the effects of changes in F0 were no longer significant. This might be attributed to the fact that English is a non-tonal language. Only the effects of changes in the spectral envelope exceeding the range from 0.88 to 1.08 for female speakers ($p < 0.05$) and from 0.92 to 1.08 for male speakers ($p < 0.01$) were significant. This suggests that clean speech recognition systems are more sensitive to shifts in the spectral envelope. This represents the first report of such comparison results.

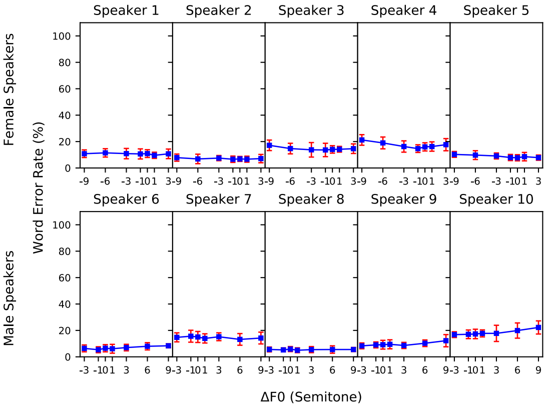

Figure 5: *WER of different speakers as a function of ΔF0 in semitones. The error bars represent the 95% confidence intervals.*

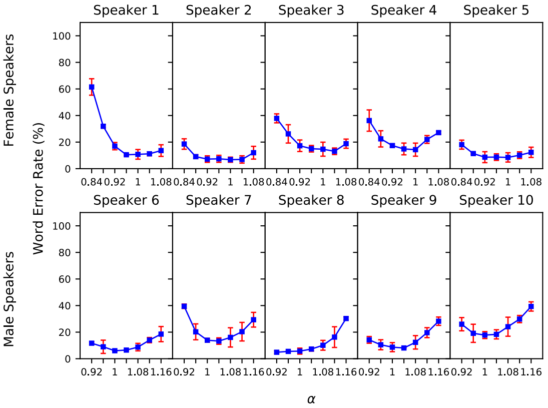

Figure 6: *WER of different speakers as a function of α. The error bars represent the 95% confidence intervals.*

3.3.2. *Evaluation of the joint speech separation and recognition system*

We connected the PIT speech separation system and the clean speech automatic recognition model to form a multi-speaker ASR system and used the joint system to perform multi-speaker speech recognition. The recognition results are plotted in FIG 7.

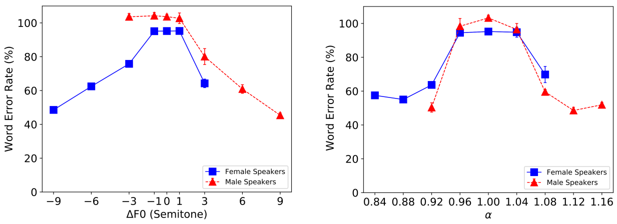

Figure 7: *The left panel shows the performance of the multi-speaker ASR system as a function of* $\Delta F0$*. The data is averaged for each gender . The right panel shows the relation between WER and* α*. The error bars represent the 95% confidence intervals.*

By comparing FIG 2, FIG 4 and FIG 7, we can see that the multi-speaker recognition performance was affected by differences in F0 or the spectral envelope and the effects on speech separation and recognition performance were statistically significant ($p < 0.001$) and consistent with each other. Results of Pearson correlation analysis on the normalized effects are shown in Table 1.

Table 1: *Correlation coefficients between speech separation and recognition results*

SPEAKER GENDER	$Female$	$Male$
SHIFTING F0	-0.898	-0.928
SHIFTING THE ENVELOPE	-0.819	-0.951

Results suggest that although small changes (when $\Delta F0 \leq$ 1 semitone or when the difference in $\alpha \leq 4\%$) do not improve the performance of a multi-speaker ASR system, larger changes do.

4. Conclusions and discussion

In [7], Darwin and his colleagues conducted experiments to explore how differences in the natural F0 or the vocal-tract length (i.e., spectral envelope) between two sentences (uttered by the same speaker) affect the human listener's ability to attend to one target sentence. They asked listeners to recognize colors and numbers in two-speaker speech mixtures following a certain 'call sign' word. Their task is highly similar to our task in Experiment 2.

In the experiment of shifting F0, they evaluated human performance averaged across SNRs of -6, -3, 0, and +3dB, which is more difficult than our setting of averaging across -2.5 to 2.5dB (Section 2.1). Their results showed that when ΔF0 was greater than 2 semitones, the listeners' performance improved significantly. More specifically, when ΔF0 was zero semitone, the recognition accuracy was about 40%; when ΔF0 equals 3 semitones, the accuracy could reach 60%; when ΔF0 equals 9 semitones, the accuracy approached 70%. For comparison, our system can achieve a WER of 50% at 9 semitones. This represents the first comparison of such kind.

When Darwin and his colleagues shifted the vocal tract length (i.e., spectral envelope) by 1.08 or higher, the change led to improved listening accuracy. In the case of -3dB and two speakers with the same envelope, the recognition accuracy was about 40%. Although the way they changed the vocal tract length was different from ours, our performance curves and theirs exhibit consistent trends. Humans performed better than automatic systems under similar experimental conditions even though they were given more difficult tasks. This means that systems based on artificial neural networks rely on F0 and the spectral envelope as humans do and the neural network models can be further improved.

For the first time, the effects of differences in F0 and the spectral envelope on deep learning-trained speech separation and recognition models were investigated in two experiments. The findings are as follows:

- The main effect of a difference in F0 on speech separation is that performance gradually improves as the difference increases, with little improvement at 1 semitone separation, which is similar to human performance. In general, speech separation systems perform better on female datasets than on male datasets.
- The performance of speech separation systems is significantly improved when two speech signals differ in their spectral envelope by a scaling factor of more than 1.08 or less than 0.92 .
- With multi-speaker automatic speech recognition models, the error mainly comes from the front-end speech separation. This suggests that the most important issue to be solved is speech separation.
- In summary, although the dependence of multi-speaker ASR systems on F0 and the envelope is similar to that of humans, machines perform worse than human auditory systems under the same conditions. This finding is new and original.

Finally, according to previous studies using human listeners [7], combined changes in F0 and the spectral envelope would produce multiplicative improvement in performance. Future studies to examine whether deep learning ASR would also exhibit similar effects or deviate from human listeners are desirable.

5. Acknowledgements

The authors would like to thank the Science, Technology and Innovation Commission of Shenzhen Municipality for partially supporting the work under project no. JCYJ20170413173515472. This study was also partially funded by the Hong Kong Research Grants Council under project no. 16200915.

6. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT press, 1994.

[3] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[4] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1670–1679, 2015.

[5] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *The Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, 1976.

[6] N. Wood and N. Cowan, "The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, no. 1, p. 255, 1995.

[7] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2913–2922, 2003.

[8] J. Kreitewolf, S. R. Mathias, R. Trapeau, J. Obleser, and M. Schönwiesner, "Perceptual grouping in the cocktail party: Contributions of voice-feature continuity," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. 2178–2188, 2018.

[9] P. Divenyi, *Speech separation by humans and machines*. Springer Science & Business Media, 2004.

[10] S. Bressler, S. Masud, H. Bharadwaj, and B. Shinn-Cunningham, "Bottom-up influences of voice continuity in focusing selective auditory attention," *Psychological research*, vol. 78, no. 3, pp. 349–360, 2014.

[11] S. R. Mathias and K. von Kriegstein, "How do we recognise who is speaking," *Front Biosci (Schol Ed)*, vol. 6, pp. 92–109, 2014.

[12] P. Divenyi, "Dimensions of auditory segregation: What do they tell us about levels of auditory processing?" pp. 468–476, 2001.

[13] J. M. Hillenbrand and M. J. Clark, "The role of f 0 and formant frequencies in distinguishing the voices of men and women," *Attention, Perception, & Psychophysics*, vol. 71, no. 5, pp. 1150–1166, 2009.

[14] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[15] Y. Qian, C. Weng, X. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[16] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[17] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Ninth International Conference on Spoken Language Processing*, 2006.

[18] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[19] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[20] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[21] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[22] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint arXiv:1704.01985*, 2017.

[23] C. Veaux, J. Yamagishi, and K. MacDonald, "Cstr vctk corpus," 2010.

[24] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[25] X. Du, M. Zhu, X. Shi, X. Zhang, W. Zhang, and J. Chen, "End-to-end model for speech enhancement by consistent spectrogram masking," *arXiv preprint arXiv:1901.00295*, 2019.

[26] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[27] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6–10.

[28] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[29] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, and T. Takano, "Real-time 2 dimensional sound source localization by 128-channel huge microphone array," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, 2004, pp. 65–70.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[31] A. Zhang, "Speech recognition." [Online]. Available: https://github.com/Uberi/speech_recognition#readme.