ON THE IMPACT OF LANGUAGE FAMILIARITY IN TALKER CHANGE DETECTION

Neeraj Sharma[1]*, Venkat Krishnamohan*[2]*, Sriram Ganapathy*[2]*, Ahana Gangopadhayay*[3]*, Lauren Fink*[4 5]

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA.
[2]Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.
[3]Electrical And Systems Engineering, Washington University in St. Louis, MO, USA.
[4]Center for Mind and Brain, Univ. of California, Davis, CA, USA.
[5]Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany.

ABSTRACT

The ability to detect talker changes when listening to conversational speech is fundamental to perception and understanding of multi-talker speech. In this paper, we propose an experimental paradigm to provide insights on the impact of language familiarity on talker change detection. Two multi-talker speech stimulus sets, one in a language familiar to the listeners (English) and the other unfamiliar (Chinese), are created. A listening test is performed in which listeners indicate the number of talkers in the presented stimuli. Analysis of human performance shows statistically significant results for: (a) lower miss (and a higher false alarm) rate in familiar versus unfamiliar language, and (b) longer response time in familiar versus unfamiliar language. These results signify a link between perception of talker attributes and language proficiency. Subsequently, a machine system is designed to perform the same task. The system makes use of the current state-of-the-art diarization approach with x-vector embeddings. A performance comparison on the same stimulus set indicates that the machine system falls short of human performance by a huge margin, for both languages.

***Index Terms*—** Talker change detection, language familiarity, benchmarking speaker diarization, response time, human versus machine.

1. INTRODUCTION

When a speech signal is processed in the human brain, we not only extract the underlying message but also extract para-linguistic attributes, such as the identity, dialect, age, and emotional state of the talker [1]. Evidence from behavioral studies suggests that perception of voice attributes influences speech processing. For example, speech in noise perception is more intelligible when a talker is familiar [2, 3, 4] and familiarity with accent impacts interpretation of the meaning of utterances [5]. Interestingly, perceptual learning of talker identity also enhances speech intelligibility in both quiet [2] and acoustically-cluttered environments [3, 4]. Further, talker-dependent adaptability in perception can be induced from exposure to just a few sentences [6]. These benefits hint at listeners' ability to track talkers in conversational speech, even in the absence of visual or spatial cues.

Detecting a change in talker identity seems to rely upon an ability to track regularities and deviations in the perceived features specific to a talker. Lavner et al. [7] suggest that talkers are identified by a distinct group of acoustic features. In a similar way, Kimberly et al. (2003) [8] have described inattention to talker changes in the context of listening for comprehension as a form of talker change deafness [9]. Further, Neuhoff et al. [10] found interactions between change detection and attention to language (and indexical) attributes. On average, talker change detection (TCD) can happen within 700 msec; this response time can be predicted from vocal tract feature difference between speech segments before and after the change instant [11, 12]. However, it is unclear if TCD is influenced by language familiarity, a deeper understanding of which could help to establish a bi-directional link between perceived phonological representation and voice perception. Evidence from the literature does suggest a dependence between linguistic proficiency and talker identification. For example, talker identification improves with increasing complexity in speech, that is, from vowels to words to sentences [13]. Also, repeated exposure to a foreign language does not suffice to improve talker identification performance in the foreign language [14]. Further, dyslexic listeners have poor talker identification even in their native language [15].

This paper aims at furthering the understanding of TCD performance when multi-talker speech utterances are from an unfamiliar language. Towards this end, an experimental setup is designed where every trial contains two sentences which have been carefully chosen not to have any contextual continuity. The sentences can be either from the same speaker or from different speakers (sex matched). The listener in the experiment is asked to report the number of talkers (1 or 2) in the trial after hearing the two sentences. The experiment contains an equal number of same and different talker trials. Using trials from a proficient language (English) and from an unknown language (Chinese), we analyze the impact of language familiarity in talker change decisions and response times.

In the design of machine systems for talker change detection, a focus on speaker diarization has attracted renewed interest [16, 17]. Most of the testing data for speaker diarization is in the English language with large amounts of in-language training data and recordings of significantly long duration ($5-10$ minute recordings). In this paper, we attempt to benchmark the diarization systems on the task done by the humans participants. Specifically, this analysis highlights machine vs. human performance for stimuli of short duration. The link for our experimental setup can be found below. [1]

This work started at the Telluride Neuromorphic Workshop in Telluride, Colorado during the summer of 2019 supported by funds from the National Science Foundation (NSF). The work done was supported by the Pratiksha Trust Young Investigator Award.

[1]https://htmlpreview.github.io/?https://github.com/iiscleap/langtcd_demo/blob/master/rt_speech.html

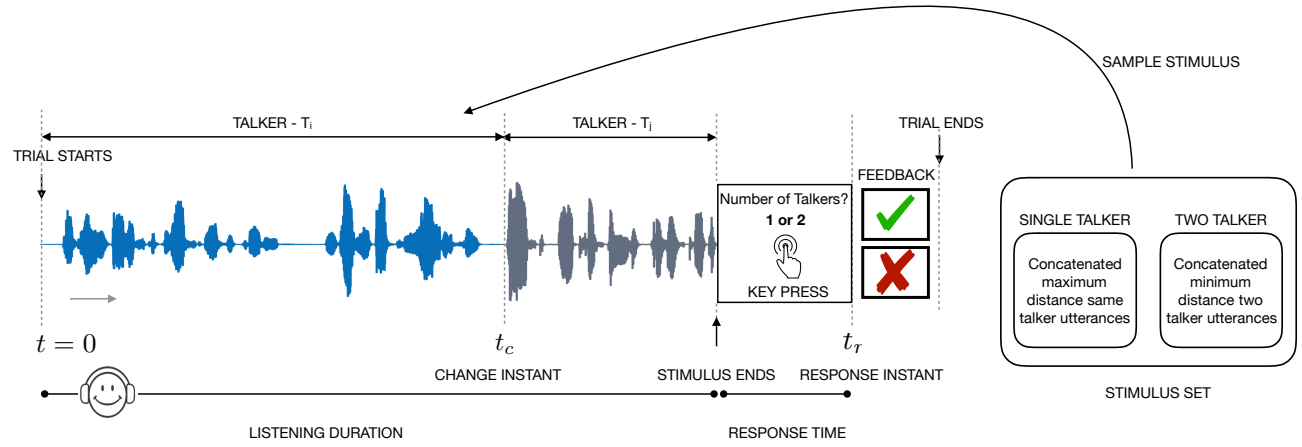

Fig. 1. Illustration of the experimental setup used in the listening test. Two sentences from either a single talker or two talkers are spliced to form the stimuli. The two sentences do not share any contextual relationships.

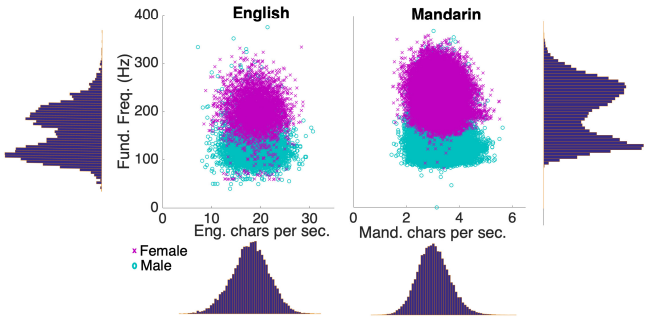

Fig. 2. Scatter plot of the stimuli features in English and Chinese. The English character units are grapheme sized English alphabets while the Chinese character units are syllable sized Chinese characters.

2. MATERIALS AND METHODS

The experimental paradigm proposed to probe the human performance for detecting multiple talkers in speech utterances is schematized in Fig. 1.

Stimuli: Each English language stimulus was composed of two utterances sourced from audiobooks taken from the LibriSpeech audiobook corpus [18], a public-domain corpus of audio data corresponding to audiobooks by 1000 different talkers. These audiobooks feature natural speech intonations. Both male and female talkers are considered, as labeled in the corpus. In two talker stimuli, the concatenation was performed only with speakers of the same sex. The utterances (single sentence level) were always drawn from different stories, or different parts of a story, so that semantic continuity did not provide a clue to talker continuity. For the Chinese stimuli, we used the Aishell corpus [19] containing 500k recordings from 400 speakers in a high fidelity, noise-free environment. In these recordings, the talking varied among five broad topics like "Finance", "Sports", "Science", "Entertainment" and "News". For both English and Chinese stimuli, we considered sentences which were between $2.5 - 5$ sec in duration.

An independent informal listening test with utterances drawn from the corpus revealed that human participants easily perceive speaking rate and average fundamental frequency differences across talkers. Fig. 2 provides an illustration of the variability along these two feature dimensions for the corpus. Here, the fundamental frequency and speaking rate were measured using the Kaldi pitch estimator [20] and the characters-per-second extracted from the spoken text, respectively. To make talker change detection challenging, we resorted to concatenating utterances based on the euclidean distance in this 2-D space. For single talker stimuli, we chose utterances from the same talker that were maximimally distant in the feature space. For multi-talker stimuli, we chose utterances from across talkers that were minimally distant in the feature space. All stimuli were manually verified to be devoid of any noise or other channel distortions.

Following this strategy, two sets of stimuli were designed (100 trials each) to focus on the language familiarity aspect of the participants. The first set was all English language utterances (known) and the second set contained all Chinese language utterances (unknown). No talker appeared in both of the sets and no stimulus was repeated within a set. On average, the duration of a single trial was $7.35(\pm 1.14)$ seconds (English) and $7.30(\pm 1.02)$ seconds (Chinese). In each stimulus set, exactly half of the trials had a single speaker while the other half had two speakers. The trials from the two languages were not mixed in the listening experiment with each stimulus set presented in its respective session. Whether participants started with the English or Chinese set was counterbalanced across participants.

Participants: A total of 14 human participants (age between $21 - 27$; mean age 24.2 and university students) with self reported normal hearing participated. All were proficient in English (confirmed by their university education curriculum) and had no prior exposure to Chinese. The protocol for the behavioral experiment was approved by the Indian Insitute of Science Ethics Board. All participants provided written consent for the test.

Procedure: Fig. 1 illustrates the listening test experiment setup. The experiment was carried out in isolated sound booth and participants listened to the stimuli through headphones. The experiment was presented using a GUI developed with Python and HTML. After presentation of each stimulus, the listeners responded with a button press indicating the number of talkers (1 or 2) in the stimulus. A visual feedback (correct/incorrect) was provided to the participant after every trial.

On average each session took 20 minutes and there was a 10 minutes break between sessions, making the total experiment

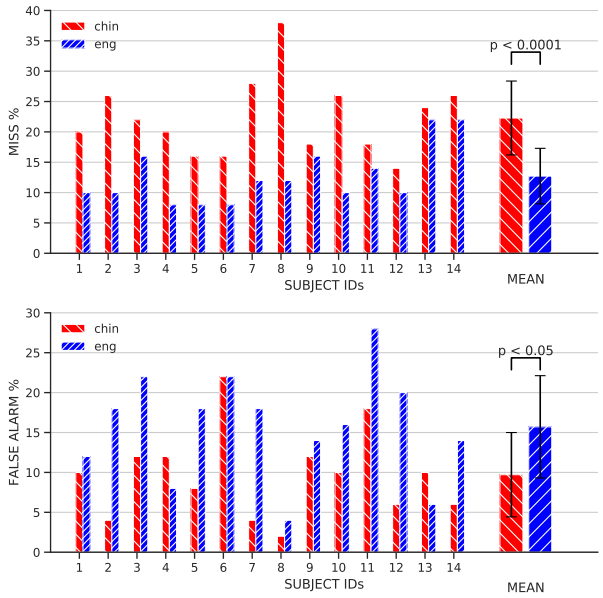

Fig. 3. Depiction of human performance on english (familiar) and chinese (unfamiliar) stimulus set as percentage of miss (*top*) and false alarm (*bottom*) responses. The error bars indicate one standard deviation.

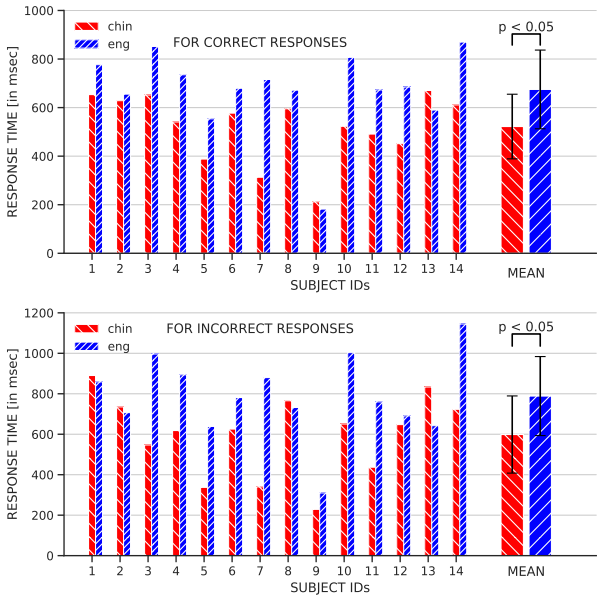

Fig. 4. Depiction of human response time on english (familiar) and chinese (unfamiliar) stimulus set. The error bars indicate one standard deviation.

duration around 45 minutes. We find that this experimental setup challenges the continuous perception faculties of the auditory system to detect, compare, and subsequently categorize the stimuli. This is unlike previous behavioral studies exploring voice perception [14, 15] which have specifically focused on talker identification after voice exposure training on each talker.

3. HUMAN PERFORMANCE RESULTS

The following measures are used to analyze the performance of the human participants.

- **Miss rate:** The percentage of two talker trials that were reported as having only one talker.
- **False alarm rate:** The percentage of single talker trials that were reported to be multi-talker trials.

The performance of the human participants in the talker detection experiment is reported in Fig. 3. As seen here, the performance is different across the two languages. In particular, the Chinese language trials showed a significantly higher miss-rate compared to the English trials, $t(26) = 4.53, p = 0.0001$. In terms of the false alarm-rate, the role of the two languages are reversed, that is, the Chinese trials had much lower false alarm rates compared to the English trials, $t(26) = -2.60, p = 0.015$. It is also interesting to note that this effect is remarkably consistent among all the participants in the listening test. These results indicate a fundamental difference in talker change detection for familiar vs. unfamiliar languages and suggest an interplay between perception of voice attributes and semantics. The results are also in agreement with talker identification studies based on exposure training [14].

4. HUMAN RESPONSE TIME

The response time in this task corresponds to the time taken after the end of the second sentence to provide the decision. The trials in which the participants took more than 2 s for a decision were considered as outliers and were excluded from the mean computation (less than 2% of trials). Further, we analyzed the response times for correct and incorrect responses separately. The mean response times across subjects is shown in Fig. 4. For most of the participants, the response time for the English task is greater than the response time for Chinese. Comparing between languages, we find a statistically significant difference for both correct ($t(26) = -2.63, p = 0.014$) and incorrect responses ($t(26) = -2.51, p = 0.019$).

The average response time over all the 14 participants in English (0.67 s) was also similar to the previous observations of response time of 0.69 s reported in [11]. However, the response time for Chinese trials was 0.15 s less than those for English. We hypothesize that this difference may occur as a result of interference, i.e. in a known language, the semantics of the spoken utterances distract from the ability to focus on talker features, increasing cognitive load and delaying the talker change decision. Such semantic interference does not occur in the case of an unknown language and results in faster response times. This notion of interference and increased cognitive load is in line with previous research, e.g. [21, 22, 23, 24] but remains to be further tested using the current paradigm.

5. MACHINE SYSTEM FOR TCD

5.1. System Description

The speaker diarization is based on using x-vector embeddings followed by a probablistic linear discriminant analysis (PLDA) approach [25]. The model is implemented in Kaldi [26].

5.1.1. Feature Extraction

The acoustic features used are 30 dimensional mel frequency cepstral coefficients (MFCC), extracted over 25 msec short-time speech segments with 10 msec overlap over temporal shifts, for training the

x-vector system [27]. A sliding mean normalization was applied over a $3s$ window.

5.1.2. *x-vector Extractor*

For training, we use a combination of VoxCeleb 1 and VoxCeleb 2 [28, 29] augmented with additive noise and reverberation according to the recipe from [30]. Segments under 4 secs in duration are discarded, resulting in a training set with $7,323$ speakers. Speech samples with reverberation are augmented by convolution of clean speech samples with room responses from the RIR dataset [31], while noisy speech augmentation is done using additive noises drawn from the MUSAN dataset [32]. The x-vector model is a time delay neural network (TDNN) [27] with 5 layers of frame level features followed by a segment level pooling of the statistics like mean and standard deviation. There are two feedforward layers following the statistical pooling layer. The final layer is the speaker target layer implementing a softmax activation and the entire model is trained on cross-entropy loss. The x-vector embeddings are the 512 dimensional hidden layer activations immediately after the segment pooling layer. At test time, x-vectors are extracted from 1.0 s segments with 0.5 s overlap over temporal shifts.

5.1.3. *PLDA training and scoring*

Probabilistic Linear Discriminant Analysis (PLDA) is used to model speaker and channel variability space. To adapt the PLDA matrix to the speaker change stimuli, a PCA transformation trained on the development dataset is applied to the training set, followed by length normalization. An utterance level PCA is applied before PLDA scoring for dimensionality reduction [33].

5.1.4. *Agglomerative Hierarchical Clustering (AHC)*

The AHC hierarchically clusters the segments based on speaker similarity scores (PLDA scores) and merges the clusters that represent the same speaker identity. Clusters are merged until their similarities dip below a stopping threshold. We vary the threshold from -0.250 to 0.250, in increments of 0.005, to obtain values for the detection error tradeoff, as plotted in Fig. 5.

5.2. Results

The performance of the speaker diarization system for English and Chinese trials is shown in Fig. 5. This plot is obtained by sweeping the threshold used in AHC clustering. The obtained miss rate and false alarm rate allow a better evaluation of the system. The human performance is the miss rate and false alarm rate computed for each of the 14 participants as well as the average over all. As seen in this plot, the human performance is significantly better than the machine performance.

In the typical evaluation of diarization systems [17], the performance is evaluated on long audio recordings of duration up to several minutes. Hence, in the current scenario, where the recordings range from $6-8$s in duration, the state-of-art diarization systems have significantly higher errors. Note that, the diarization outputs are only analyzed in terms of the number of speakers and not the diarization error rate (DER) metric. Even with this simplified metric, this evaluation shows that the human performance on multi-talker detection tasks has less than half the number of errors generated by a machine system. The results show that machine systems rely on large amounts of within speaker audio to perform speaker clustering. With only a small number of within speaker x-vector embeddings in the test data ($6-10$ embeddings from each sentence), the AHC algorithm has substantial trouble in identfying speaker clusters. The performance gap highlights that understanding human processing of talker change detection in short duration recordings can provide important cues for the design of improved speaker diarization systems.

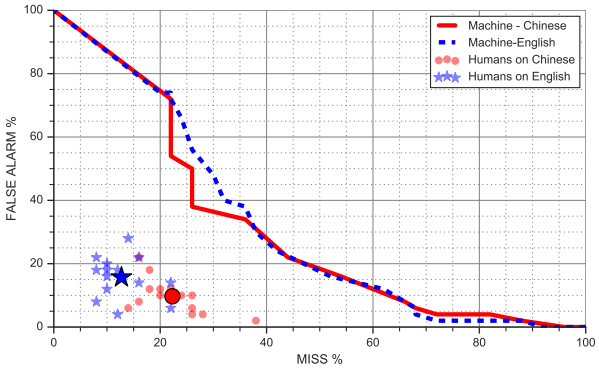

Fig. 5. Comparison of talker change detection performance for human participants with state-of-art diarization system. The diarization system output is only used for counting the number of speaker clusters.The human evaluation results in terms of individual listener miss and false-alarm rates are provided along with the mean values.

6. CONCLUSION

In this paper, we present a novel paradigm to probe the impact of language familiarity in talker change detection. We find that human human detection of talker change is impacted by language familiarity. In a known (English) vs. unknown (Chinese) language, humans have significantly higher false-alarm rates, lower miss rates, and longer response times. Compared to a state-of-the-art machine system with x-vector PLDA scoring and agglomerative hierachical clustering, human performance for both the familiar and unfamiliar languages on the talker change detection task results in fewer errors.

7. REFERENCES

[1] John D. M. Laver, "Voice quality and indexical information," *British Journal of Disorders of Communication*, vol. 3, no. 1, pp. 43–54, 1968.

[2] Lynne C. Nygaard and David B. Pisoni, "Talker-specific learning in speech perception," *Perception & Psychophysics*, vol. 60, no. 3, pp. 355–376, Jan 1998.

[3] Pádraig T. Kitterick, Peter J. Bailey, and A. Quentin Summerfield, "Benefits of knowing who, where, and when in multi-talker listening," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2498–2508, 2010.

[4] Ingrid S. Johnsrude, Allison Mackey, Hélène Hakyemez, Elizabeth Alexander, Heather P. Trang, and Robert P. Carlyon, "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," *Psychological Science*, vol. 24, no. 10, pp. 1995–2004, 2013.

[5] Zhenguang G. Cai, Rebecca A. Gilbert, Matthew H. Davis, M. Gareth Gaskell, Lauren Farrar, Sarah Adler, and Jennifer M.

Rodd, "Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition," *Cognitive Psychology*, vol. 98, pp. 73 – 101, 2017.

[6] Matthias J. Sjerps, Holger Mitterer, and James M. McQueen, "Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics," *Neuropsychologia*, vol. 49, no. 14, pp. 3831 – 3846, 2011.

[7] Yizhar Lavner, Isak Gath, and Judith Rosenhouse, "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," *Speech Communication*, vol. 30, no. 1, pp. 9 – 26, 2000.

[8] Kimberly M. Fenn, Hadas Shintel, Alexandra S. Atkins, Jeremy I. Skipper, Veronica C. Bond, and Howard C. Nusbaum, "When less is heard than meets the ear: Change deafness in a telephone conversation," *Quarterly Journal of Experimental Psychology*, vol. 64, no. 7, pp. 1442–1456, 2011.

[9] Michael S Vitevitch, "Change deafness: The inability to detect changes between two voices.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, no. 2, pp. 333, 2003.

[10] John G Neuhoff, Steven A Schott, Adam J Kropf, and Emily M Neuhoff, "Familiarity, expertise, and change detection: Change deafness is worse in your native language," *Perception*, vol. 43, no. 2-3, pp. 219–222, 2014.

[11] Neeraj Kumar Sharma, Shobhana Ganesh, Sriram Ganapathy, and Lori L. Holt, "Talker change detection: A comparison of human and machine performance," *The Journal of the Acoustical Society of America*, Oct. 2018.

[12] N. Sharma, S. Ganesh, S. Ganapathy, and L. L. Holt, "Analyzing human reaction time for talker change detection," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2019, pp. 7135–7139.

[13] Judith P. Goggin, Charles P. Thompson, Gerhard Strube, and Liza R. Simental, "The role of language familiarity in voice identification," *Memory & Cognition*, vol. 19, no. 5, pp. 448–458, Sep 1991.

[14] Tyler K. Perrachione and Patrick C.M. Wong, "Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex," *Neuropsychologia*, vol. 45, no. 8, pp. 1899 – 1910, 2007.

[15] Tyler K. Perrachione, Stephanie N. Del Tufo, and John D. E. Gabrieli, "Human voice recognition depends on language ability," *Science*, vol. 333, no. 6042, pp. 595–595, 2011.

[16] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "First DIHARD challenge evaluation plan," Tech. Rep., 2018.

[17] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "Second DIHARD challenge evaluation plan," Tech. Rep., 2019.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP)*, April 2015, pp. 5206–5210.

[19] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[20] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[21] Sarah C Creel and Melanie A Tumlin, "On-line acoustic and semantic interpretation of talker information," *Journal of Memory and Language*, vol. 65, no. 3, pp. 264–285, 2011.

[22] Thomas Koelewijn, Adriana A Zekveld, Joost M Festen, Jerker Rönnberg, and Sophia E Kramer, "Processing load induced by informational masking is related to linguistic abilities," *International journal of otolaryngology*, vol. 2012, 2012.

[23] Marie Dekerle, Véronique Boulenger, Michel Hoen, and Fanny Meunier, "Multi-talker background and semantic priming effect," *Frontiers in human neuroscience*, vol. 8, pp. 878, 2014.

[24] Chandan R Narayan, Lorinda Mak, and Ellen Bialystok, "Words get in the way: Linguistic effects on talker discrimination," *Cognitive science*, vol. 41, no. 5, pp. 1361–1376, 2017.

[25] Prachi Singh, Harsha Vardhan, Sriram Ganapathy, and Ahilan Kanagasundaram, "LEAP diarization system for the second dihard challenge," in *Interspeech*. 2019, International Speech Communication Association (ISCA).

[26] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," Tech. Rep., IEEE Signal Processing Society, 2011.

[27] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[28] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[29] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.

[30] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop*, 2016, pp. 165–170.

[31] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[32] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[33] Weizhong Zhu and Jason Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016.