Acoustic and linguistic features influence talker change detection

Neeraj Kumar Sharma, Venkat Krishnamohan, Sriram Ganapathy, Ahana Gangopadhayay, and Lauren Fink

Citation: The Journal of the Acoustical Society of America **148**, EL414 (2020); doi: 10.1121/10.0002462
View online: https://doi.org/10.1121/10.0002462
View Table of Contents: https://asa.scitation.org/toc/jas/148/5
Published by the Acoustical Society of America

JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0002462

Acoustic and linguistic features influence talker change detection

Neeraj Kumar Sharma,[1,a)] Venkat Krishnamohan,[1] Sriram Ganapathy,[1] Ahana Gangopadhayay,[2] and Lauren Fink[3,b)]

[1]*Learning and Extraction of Acoustic Patterns Lab, Indian Institute of Science, Bangalore*
[2]*Electrical and Systems Engineering, Washington University in St. Louis, Missouri, USA*
[3]*Music Department, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany*
neerajww@gmail.com, venkat201097@gmail.com, sriram.iisc@gmail.com, ahana@wustl.edu, lauren.fink@ae.mpg.de

Abstract: A listening test is proposed in which human participants detect talker changes in two natural, multi-talker speech stimuli sets—a familiar language (English) and an unfamiliar language (Chinese). Miss rate, false-alarm rate, and response times (RT) showed a significant dependence on language familiarity. Linear regression modeling of RTs using diverse acoustic features derived from the stimuli showed recruitment of a pool of acoustic features for the talker change detection task. Further, benchmarking the same task against the state-of-the-art machine diarization system showed that the machine system achieves human parity for the familiar language but not for the unfamiliar language.
 https://doi.org/10.1121/10.0002462

[Editor: Douglas D O'Shaughnessy] Pages: EL414–EL419

Received: 31 May 2020 **Accepted:** 20 October 2020 **Published Online:** 25 November 2020

1. Introduction

The perception (and decoding) of talker attributes is essential while listening to multi-talker speech conversations. In this paper, we present an experimental paradigm to probe talker change detection in human listeners with stimuli drawn from familiar and unfamiliar languages and find that change detection is dependent on language familiarity and specific acoustic features. A human-machine comparison using a diarization system shows that the performance of the machine system is on par with the human performance for the familiar language.

Previous behavioral studies suggest a substantial influence of indexical attributes, such as talker identity, dialect, age, etc. (Laver, 1968), on speech intelligibility. For example, talker familiarity improves speech in noise perception (Johnsrude *et al.*, 2013; Kitterick *et al.*, 2010; Nygaard and Pisoni, 1998) and accent familiarity alters the perceived meaning of an utterance (Cai *et al.*, 2017). These imply perception of talker cues helps in parsing the semantic message. Lavner *et al.* (2000) suggest that talker identification uses a distinct group of acoustic features. Yet, Sell *et al.* (2015) argue that a combination of vocal source, vocal tract, and cortical features fail to explain the perceived talker discrimination in a listening test with simple word-level utterances. Talker perception improves with increase in phonetic content in the speech signal, that is, from vowels to words and sentences (Goggin *et al.*, 1991). Deafness to talker change (Neuhoff *et al.*, 2014), as well as perceptual sensitivity in judging talker dissimilarity (Fleming *et al.*, 2014; Perrachione *et al.*, 2011; Perrachione *et al.*, 2019), are both found to be affected by linguistic familiarity as well. These studies suggest an interplay between phonetic, semantic, and talker perception while listening to speech.

Unlike single-talker speech, multi-talker conversations contain talker change instances and detecting these instances is required for segregating the speech into time segments corresponding to who spoke what, and when. Human listeners, on average, take approximately 700 ms (from the instant of change) to report a talker change (Sharma *et al.*, 2019). While acoustic features before and after the change instant influence change detection, it is not clear if semantic processing in a familiar language impacts talker change detection (TCD). Hence, this paper compares talker change detection using stimuli from familiar and unfamiliar languages.

We designed two speech stimuli sets, one in a language familiar to the participants (English) and another in an unfamiliar language (Mandarin Chinese, henceforth referred to as Chinese). We assume that, compared to a familiar language, semantic processing is minimal

[a)]Author to whom correspondence should be addressed.
[b)]Also at Center for Language, Music and Emotion, New York University, Max Planck Institute, NY, USA.

https://doi.org/10.1121/10.0002462

while listening to the unfamiliar language. The participants took part in a listening test to indicate the number of talkers in multi-talker stimuli derived from these datasets. The collected data were analyzed to understand the impact of language familiarity on detection metrics, namely, miss and false alarm rates, and on the use of acoustic features in responding to the task via regression modeling of the response times (RT). Further, talker change detection is identified as a crucial pre-processing step (Ryant *et al.*, 2018; Ryant *et al.*, 2019) for machine recognition of conversational speech. This step is primarily approached using diarization systems. We investigate the performance of the state-of-art diarization system based on x-vector embeddings (Snyder *et al.*, 2018) on the stimuli sets used in the human listening task. In the recent years there have been claims on achieving human parity in applications like automatic speech recognition (ASR) (Saon *et al.*, 2017; Xiong *et al.*, 2016) and machine translation (Hassan *et al.*, 2018). In this context, highlighting the performance gap, if any, between humans and machines constitutes an important step to achieve human parity for speaker diarization systems.

2. Methods

The study presented here is an extension of our work in (Sharma *et al.*, 2020a) with a larger set of human participants, and a detailed analysis of reaction time modeling.

2.1 Participants

A total of 28 human participants (21 male, age range 20–37; mean age 24 years, with self-reported normal hearing) participated in the listening test. All participants were proficient in English and had no prior exposure to Chinese. The protocol for the behavioral experiment was approved by the Indian Institute of Science Human Ethics Committee. All participants provided written consent for the test and were provided with monetary compensation.

2.2 Stimuli

The English and Chinese speech signal recordings were taken from the LibriSpeech corpus (Panayotov *et al.*, 2015) and the Aishell corpus (Bu *et al.*, 2017), respectively. These corpora are composed of read speech audio data (audiobooks and news broadcasts) from more than 400 talkers and are freely available in the public-domain. For our experiment, the single talker stimuli were formed by concatenating two utterances from the same talker, while the two-talker stimuli were formed by concatenating two utterances from two different, gender-matched talkers. Both utterances were chosen to avoid any contextual continuity, and had a duration ranging from 2.5 to 5 s, forming a stimulus of 5–10 s. With this approach, two curated stimuli sets were constructed—one for English and one for Chinese, each with 50 single talker and 50 two-talker stimuli. All the stimuli were manually checked for quality (absence of noise/channel distortions). In order to avoid any talker adaptation during listening to these stimuli, none of the talkers appeared in more than one stimulus. A comparison of the distribution of a few of the acoustic features for the stimuli in the two stimuli sets is shown in Fig. 1(b). The acoustic features, namely, pitch, harmonic-to-noise ratio (correlated with perceived voice quality), and intensity (correlated with perceived loudness), are obtained from short-time 40 ms speech segments (with temporal hop of 10 ms) derived from the speech signals [extracted using PRAAT (Boersma and Weenink, 2020)]. There is considerable overlap between the distributions, illustrating the acoustic feature similarity between the two stimuli sets. The bimodal distribution in pitch is due to male and female utterances in the stimuli sets.

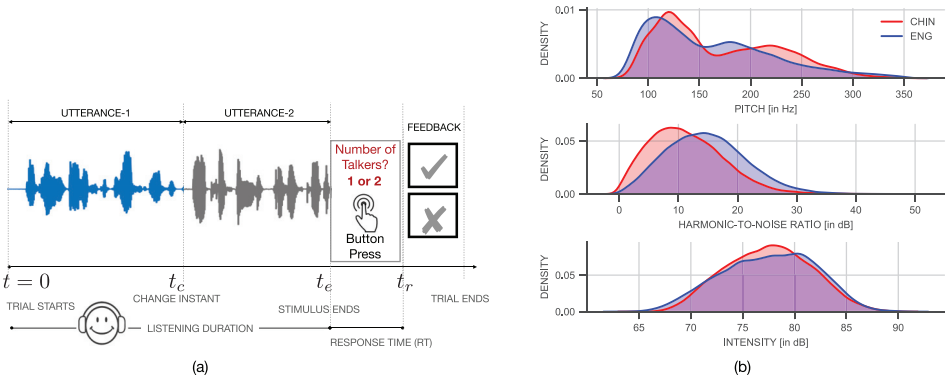

Fig. 1. (Color online) (a) Illustration of a listening test trial. (b) A comparison acoustic feature distributions between English (ENG) and Chinese (CHIN) audio stimuli sets.

2.3 Listening test

The listening test for each participant was conducted in two sessions. Each session had stimuli only from one language. The ordering of language presentations was randomized across participants, and the order of stimuli presentation in each language session was randomized for every participant. The experiment was conducted in an isolated sound booth using high fidelity headphones (Sensheiser HD 215). A graphical user interface designed in python and HTML was used for stimuli presentation and recording responses [stimulus material available at Sharma *et al.* (2020b)]. After presentation of a stimulus, the listener responded with a button press indicating the number of talkers (1 or 2). Visual feedback (correct/incorrect) was provided to the participant after every trial. An illustration of a trial is shown in Fig. 1(a). A small set of 16 trials were provided to get familiar with the task. On average, the session for each language took 20 min and there was a 10 min break between sessions, making the total experiment duration 50 min per participant.

2.4 Behavioral data pre-processing

The performance measures used are: (i) Miss rate (%): the percentage of two talker stimuli reported by the participant as single talker, (ii) False Alarms (FA) rate (%): the percentage of single talker stimuli reported as two-talker, and (iii) Response time (RT): the time duration between the end of the stimulus and the participant's response in the form of button press [that is, RT=$t_r - t_e$ shown in Fig. 1(a)]. Any trial with a response time $RT < 20$ ms (too fast) or $RT > 2$ s (too slow) was discarded for the analysis. The discarded trials constituted 6.7% of the collected responses.

2.5 Machine system

We used an implementation of a state-of-the-art speech diarization system which uses x-vector embeddings as acoustic features. The x-vector embeddings from short speech segments are fed to a probabilistic linear discriminant analysis (PLDA) to generate the affinity matrix. The PLDA affinity matrix is used by an agglomerative hierarchical clustering (AHC) framework to cluster x-vector features. The output is talker-level segmentation of the input speech signal. We consider the system output hypothesis as two talkers if more than one talker is present in the segmentation. The system implementation details are provided in Singh *et al.* (2019). The x-vector embeddings (Singh *et al.*, 2019) are derived from a hidden layer of a time-delay neural network trained for a talker classification task on the VoxCeleb-1 and VoxCeleb-2 [celebrity speech corpus (Chung *et al.*, 2018) composed of 7323 talkers]. These embeddings (512 dimensional) capture the talker attributes derived from 1 s segments of speech. The threshold for the AHC clustering was varied from –0.250 to 0.250, in increments of 0.005, to compute the miss and false-alarm probabilities. These values were used to obtain the detection error trade-off curve plotted in Fig. 3(d).

3. Results

3.1 Behavioral data

A scatter plot of miss-rate and FA-rate for unfamiliar (Chinese) versus familiar language (English) stimuli sets is shown in Figs. 2(a) and 2(b). A majority of the participants showed a higher miss-rate for Chinese trials and a higher FA-rate for English trials. The d-prime for the task [Fig. 2(c)] was found to be greater than 1.5 for most of the participants indicating the participants performed the task effectively. Also, the bias [Fig. 2(d)] was between 0.4 and 3 with a higher spread for Chinese trials. The miss and FA averaged across participants is shown in Figs. 2(e) and 2(f). The average miss-rate is significantly higher for the unfamiliar language [that is, Chinese, with $t(56) = 2.38$, $p < 0.05$, Cohen's $-d = 0.64$]. The average FA-rate is significantly higher for the familiar language [that is, English, with $t(56) = -2.80$, $p < 0.01$, Cohen's $-d = -0.74$]. The distributions of pooled RTs (from all participants) for correct and incorrect responses are shown in Fig. 2(g); these are visually distinct for the two languages. The grand average of participants' mean RT is shown in Figs. 2(h) and 2(i). The average RT for unfamiliar language (Chinese) is significantly smaller [with $t(56) = -3.02$, $p < 0.005$, Cohen's $-d = -0.81$ for correct responses, and $t(56) = -4.09$, $p < 0.005$, Cohen's $-d = -1.09$ for incorrect responses]. These observations indicate a significant impact of language familiarity on human TCD performance.

3.2 Linear regression modeling of RTs

A linear regression model was constructed with acoustic feature distances as predictor variables and the RT as the dependent variable. As RT is always greater than zero and has a skewed distribution [see Figs. 2(e), 2(f)], the natural logarithmic transformation of RT was used. The acoustic features included: mel-spectrogram (MEL; using 40 filters), mel-frequency cepstral coefficients

JASA EXPRESS LETTERS
https://doi.org/10.1121/10.0002462

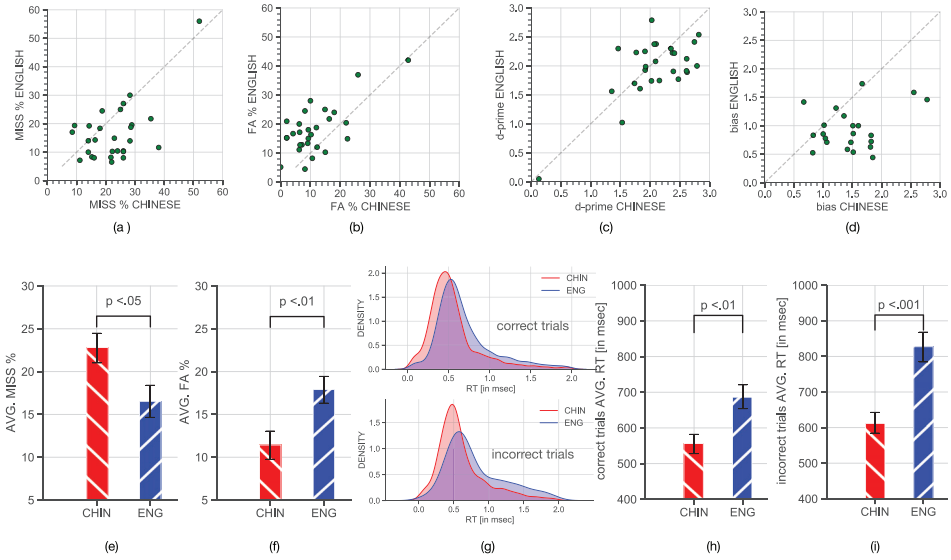

Fig. 2. (Color online) Human performance on the talker change detection task, as a function of language familiarity. (a), (b) miss and false alarm rates, respectively, for each participant; (c), (d) d-prime and bias for each participant; (e), (f) average miss and false rates; (g) all participants' pooled response times on correct (top), and incorrect trials (bottom); (h), (i) average response times on correct, and incorrect trials, respectively. All error bars represent the standard error of the mean.

(MFCC; 13 coefficients), intensity (INTENSITY), spectral centroid (SCENTROID), pitch (PITCH), harmonic-to-noise ratio (HNR), and x-vectors (XVEC, features used in the machine system). Given a stimulus signal, for each feature type, we obtain two representations - one for each of the concatenated utterances. These feature representations correspond to average of short-time frame-wise (40 ms, with temporal hop of 10 ms; unvoiced frames were discarded) extracted features [using PRAAT; McFee *et al.* (2020)]. The feature distance is measured as the Euclidean distance between the mean of feature representations from the two utterances. Alongside the acoustic features distances, we also included stimulus duration (T_d) as a predictor variable. As there is a significant impact of language type on RT (seen in Sec. 3.1), we model RTs separately for different subsets of the pooled data. We have eight models basing on language (Chinese/English), response (correct/incorrect trials), and trial stimulus type (two talker/single talker). Figure 3(a) shows the result obtained from a type-II analysis of variance (ANOVA) on every model. There is variability in the RTs across subjects making the subject identity (SUB_ID), a categorical predictor variable, significant in all the models. With respect to acoustic features, more acoustic features are significant for English compared to Chinese stimuli. The R^2 is also high for English compared to Chinese implying a relatively higher percentage of the

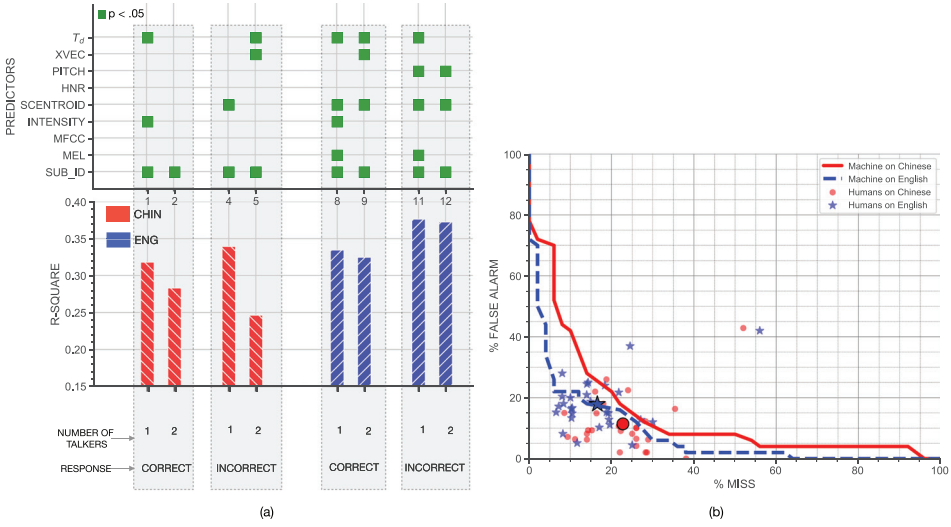

Fig. 3. (Color online) (a) Top: Feature significance across models, green square indicates $p < 0.05$. Bottom: Model R^2. (b) Detection error trade-off (DET) curve for the machine system. The scattered data points correspond to human participants. The two highlighted larger data points correspond to average across human participants.

observed data variance explained by the predictors for English stimuli. Interestingly, the stimulus duration is also found to be of significance in most of the models. Surprisingly, MFCC and HNR did not turn out to be of significance in any model and SCENTROID was significant in the majority of the models. The XVEC was found to be significant for two-talker correct English trials. This is interesting as the x-vector features are designed to capture talker differences and have been shown to be useful in machine diarization systems.

3.3 Human-machine comparison

The machine system performance is shown in Fig. 3(b). The human performance is also included in this figure. The plot suggests that performance on the familiar language (English) for the machine system is on par with human performance for the same stimuli. The performance on the unfamiliar language (Chinese) is worse for the machines, compared to human performance. This indicates that the future design of machine diarization systems could target invariance to language mis-match.

4. Discussion

The listening test results from the study show a significant impact of language familiarity on human talker change detection performance. Though the sound stimuli we used were short in duration (2.5–5 s utterances), each utterance had close to 8–10 words and hence, was not devoid of semantic information. Such short duration, sentence-level speech stimuli have previously been used for analyzing language familiarity effects on speaker dissimilarity judgments by Perrachione *et al.* (Perrachione *et al.*, 2019) and Flemming *et al.* (Fleming *et al.*, 2014). These studies highlight that even using time-reversed speech devoid of semantics is sufficient to illustrate a familiarity effect.

The results show a lower miss rate for familiar language suggesting that success in semantic processing (and understanding) benefits TCD. However, we also find that the FA is higher for the familiar language. This suggests that a majority of participants falsely associated a change in context between the utterances with a talker change. This is not the case for the unfamiliar language (significantly lower FA) as the semantic understanding is absent. The RT for familiar language trials is significantly higher compared to the unfamiliar language trials. This finding suggests that comprehension of speech (which likely occurs in familiar language stimuli) adversely affects the TCD response time, whereas in the unfamiliar language case, there is no conflict (increased cognitive load) of semantic processing involved. Past work by Neuhoff *et al.* (Neuhoff *et al.*, 2014) presents an interesting interplay between semantic and indexical information extraction, showing greater change deafness for familiar language (without participants being cued to attend to the change). In contrast to that study, the subjects in the current study were instructed to attend to talker changes and were also provided feedback after every trial. Therefore, even when we instruct participants to attend to the change, we still find effects of language familiarity on change detection. In particular, listening to familiar language speech distracts from the ability to attend to indexical information, which likely manifests as the increased reaction times observed in the familiar language trials.

We note that the subject pool recruited for the study consisted of non-native English speakers, proficient in English. Data from our past study (Sharma *et al.*, 2020) and also from Köster and Schiller (1997) suggests non-nativeness does not have an impact on talker perception tasks, though future studies may wish to manipulate this factor.

Moving to the regression analysis of RTs, we find that a majority of the acoustic features failed to be of significance for the unfamiliar language trials. This was also reflected in a lower R^2 for the data drawn from trials corresponding to the unfamiliar language. We hypothesize that language familiarity enables usage of acoustic features which are different from those used for unfamiliar language.

To the best of our knowledge, this study is the first of its kind to contrast human and machine performance on a talker counting task. The human-machine performance comparison shows that the diarization systems based on x-vector embeddings can achieve human-like performance even on short duration stimuli when the training and test data come from the same language. However, the results indicate that humans are superior in generalizing to unfamiliar languages. The future design of embeddings for diarization systems can target language invariance to overcome this limitation.

Acknowledgments

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 by N.K.S., V.K., and S.G. was supported by grants from the British Telecom India Research Center (BTIRC).

https://doi.org/10.1121/10.0002462

References and links

Boersma, P., and Weenink, D. (**2020**). "Praat: Doing phonetics by computer," www.praat.org (Last viewed November 3, 2020).

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. (**2017**). "Aishell-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *IEEE Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5.

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., and Rodd, J. M. (**2017**). "Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition," Cogn. Psychol. **98**, 73–101.

Chung, J. S., Nagrani, A., and Zisserman, A. (**2018**). "VoxCeleb2: Deep speaker recognition," in *Proceedings of Interspeech*, pp. 1086–1090.

Fleming, D., Giordano, B. L., Caldara, R., and Belin, P. (**2014**). "A language-familiarity effect for speaker discrimination without comprehension," Proc. Natl. Acad. Sci. **111**(38), 13795–13798.

Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (**1991**). "The role of language familiarity in voice identification," Mem. Cogn. **19**(5), 448–458.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (**2018**). "Achieving human parity on automatic Chinese to English news translation," arXiv:1803.05567.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**(10), 1995–2004.

Kitterick, P. T., Bailey, P. J., and Summerfield, A. Q. (**2010**). "Benefits of knowing who, where, and when in multi-talker listening," J. Acoust. Soc. Am. **127**(4), 2498–2508.

Köster, O., and Schiller, N. O. (**1997**). "Different influences of the native language of a listener on speaker recognition," Foren. Ling. **4**, 18–28.

Laver, J. D. M. (**1968**). "Voice quality and indexical information," Brit. J. Disord. Commun. **3**(1), 43–54.

Lavner, Y., Gath, I., and Rosenhouse, J. (**2000**). "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," Speech Commun. **30**(1), 9–26.

McFee, B., Lostanlen, V., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Mason, J., Ellis, D., Battenberg, E., Seyfarth, S., Yamamoto, R., Choi, K., viktorandreevichmorozov, Moore, J., Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F.-R., Friesch, P., Weiss, A., Vollrath, M., and Kim, T. (**2020**). "librosa/librosa: 0.8.0," https://doi.org/10.5281/zenodo.3955228 (Last viewed November 3, 2020).

Neuhoff, J. G., Schott, S. A., Kropf, A. J., and Neuhoff, E. M. (**2014**). "Familiarity, expertise, and change detection: Change deafness is worse in your native language," Perception **43**(2-3), 219–222.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Perception Psychophys. **60**(3), 355–376.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processes (ICASSP)*, pp. 5206–5210.

Perrachione, T. K., Del Tufo, S. N., and Gabrieli, J. D. E. (**2011**). "Human voice recognition depends on language ability," Science **333**(6042), 595–595.

Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (**2019**). "Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices," J. Acoust. Soc. Am. **146**(5), 3384–3399.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (**2018**). "First DIHARD challenge evaluation plan," technical report, https://catalog.ldc.upenn.edu/docs/LDC2019S09/first_dihard_eval_plan_v1.3.pdf (Last viewed November 3, 2020).

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (**2019**). "The second dihard diarization challenge: Dataset, task, and baselines," in *Proceedings of Interspeech*, pp. 978–982.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., and Hall, P. (**2017**). "English conversational telephone speech recognition by humans and machines," arXiv:1703.02136.

Sell, G., Suied, C., Elhilali, M., and Shamma, S. (**2015**). "Perceptual susceptibility to acoustic manipulations in speaker discrimination," J. Acoust. Soc. Am. **137**(2), 911–922.

Sharma, N., Krishnamohan, V., Ganapathy, S., Gangopadhayay, A., and Fink, L. (**2020a**). "On the impact of language familiarity in talker change detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6249–6253.

Sharma, N., Krishnamohan, V., Ganapathy, S., Gangopadhayay, A., and Fink, L. (**2020b**). *Resources for impact of language on talker change detection task*, www.github.com/iiscleap/langtcd_demo (Last viewed April 24, 2020).

Sharma, N. K., Ganesh, S., Ganapathy, S., and Holt, L. L. (**2019**). "Talker change detection: A comparison of human and machine performance," J. Acoust. Soc. Am. **145**(1), 131–142.

Singh, P., Vardhan, H., Ganapathy, S., and Kanagasundaram, A. (**2019**). "LEAP diarization system for the second DIHARD challenge," in *Proceedings of Interspeech*, pp. 983–987.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (**2018**). "X-vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, pp. 5329–5333.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (**2016**). "Achieving human parity in conversational speech recognition," arXiv:1610.05256.