The Journal of the Acoustical Society of America Express Letters

# Acoustic and linguistic features influence talker change detection

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | JASA-EL-01551 |
| **Full Title:** | Acoustic and linguistic features influence talker change detection |
| **Short Title:** | Talker change detection |
| **Article Type:** | Express Letter |
| **Section/Category:** | Speech Communication |
| **Keywords:** | talker change, language familiarity, diarization |
| **Abstract:** | A listening test is proposed in which human participants detect talker changes in two natural, multi-talker speech stimulus sets - one in a familiar language (English) and the other in an unfamiliar language (Chinese). In the experiments, the miss rate, false alarm rate, and response time (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. In addition, benchmarking the same task against the state-of-the-art machine diarization system showed the human performance to be substantially better than the machine performance. |

# Acoustic and linguistic features influence talker change detection

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

[1)] *Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science, Bangalore*

[2)] *Electrical And Systems Engineering, Washington University in St. Louis, MO, USA*

[3)] *Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany*[a)]

*neerajww@gmail.com,*

*venkat201097@gmail.com,*

*sriram.iisc@gmail.com,*

*ahana@wustl.edu,*

*lauren.fink@ae.mpg.de*

(Dated: 31 May 2020)

**Abstract:** A listening test is proposed in which human participants detect talker changes in two natural, multi-talker speech stimulus sets - one in a familiar language (English) and the other in an unfamiliar language (Chinese). In the experiments, the miss rate, false alarm rate, and response time (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. In addition, benchmarking the same task against the state-of-the-art machine diarization system showed the human performance to be substantially better than the machine performance.



[a] Author to whom correspondence should be addressed.

## 1. Introduction

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). This implies perception of talker cues helps in parsing the semantic message. Lavner et al. (Lavner *et al.*, 2000) suggest that talker identification uses a distinct group of acoustic features. Yet, Sell et al. (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). Perceptual sensitivity in judging talker dissimilarity is found to be affected by linguistic familiarity (Perrachione *et al.*, 2011, 2019) as well. These studies suggest an interplay between semantic and talker processing while listening to speech.

The perception (and decoding) of talker attributes is even more essential while listening to multi-talker speech conversations. 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 msec (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 impacts talker change detection (TCD). This paper attempts to get a deeper understanding on this aspect.

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 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 time (RT). Further, talker change detection is identified as a crucial pre-processing step (Ryant *et al.*, 2018, 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. The machine results are benchmarked with the results from the human experiments. The study presented here is an extension of our work in (Sharma *et al.*, 2020) with a larger set of human participants, and a detailed analysis of reaction time modeling.

## 2. Methods

### *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 Insitute 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 $2.5 - 5$ 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 stimuli. A comparison of the distribution of few of the acoustic features for the stimuli in the two stimuli sets is shown in Figure 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 msec speech segments (with temporal hop of 10 msec) derived from the speech signals (extracted using PRAAT (Boersma and Weenink)). There is considerable

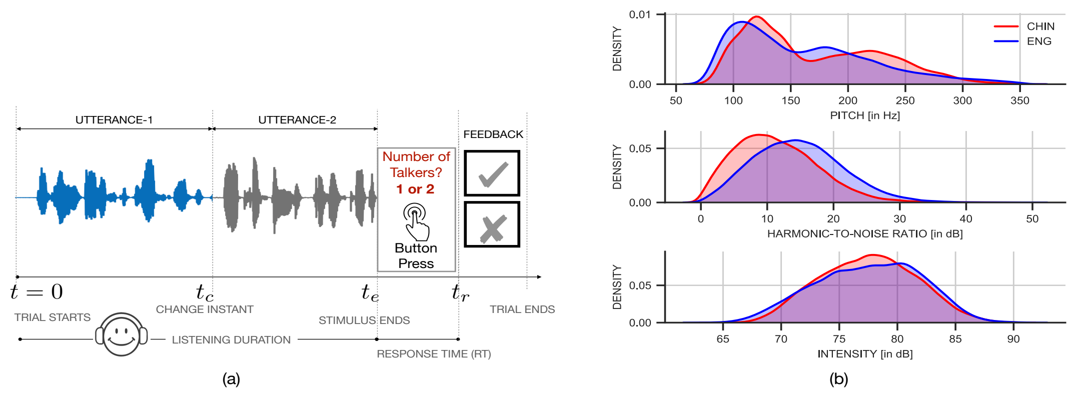


Fig. 1. (color online) (a) Illustaration of a listening test trial. (b) A comparison of distribution of three acoustic features between English (ENG) and Chinese (CHIN) audio stimuli sets.

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.

### *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. 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 (web-demo available at (Krishnamohan, 2020 (accessed April 24, 2020)). 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 Figure 1(a). On average, the session for each language took 20 mins and there was a 10 mins break between sessions, making the total experiment duration to 50 mins 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 Figure 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 probilistic 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 2 talkers if more than one talker is present in the segmentation. The system implementation details are provided in (Singh *et al.*, 2019). The x-vectors 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 sec 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 probablities. These values were used to obtain the detection error trade-off curve plotted in Figure 3(d).

## 3. Results

### *3.1 Behavioral data*

A scatter plot of miss-rate and FA-rate for unfamiliar (Chinese) versus familiar (English) stimuli sets is shown in Figure 2(a,b). A majority of the participants showed a higher miss-rate for Chinese trials and a higher FA-rate for English trials. The average across participants is shown in Figure 2(c,d). The average miss-rate is significantly higher for the unfamiliar language (that is, Chinese, with $t(56) = 2.38,\ p < .05$). The average FA-rate is significantly higher for the familiar language (that is, English, with $t(56) = -2.80,\ p < .01$). The distributions of pooled RTs (from all participants) for correct and incorrect responses are shown in Figure 2(e,f); these are visually distinct for the two languages. The grand average of participants' mean RT is shown in Figure 2(g,h). The average RT for unfamiliar language (Chinese) is significantly smaller (with $t(56) = -3.02,\ p < .005$ for correct responses, and $t(56) = -4.09,\ p < .005$ 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

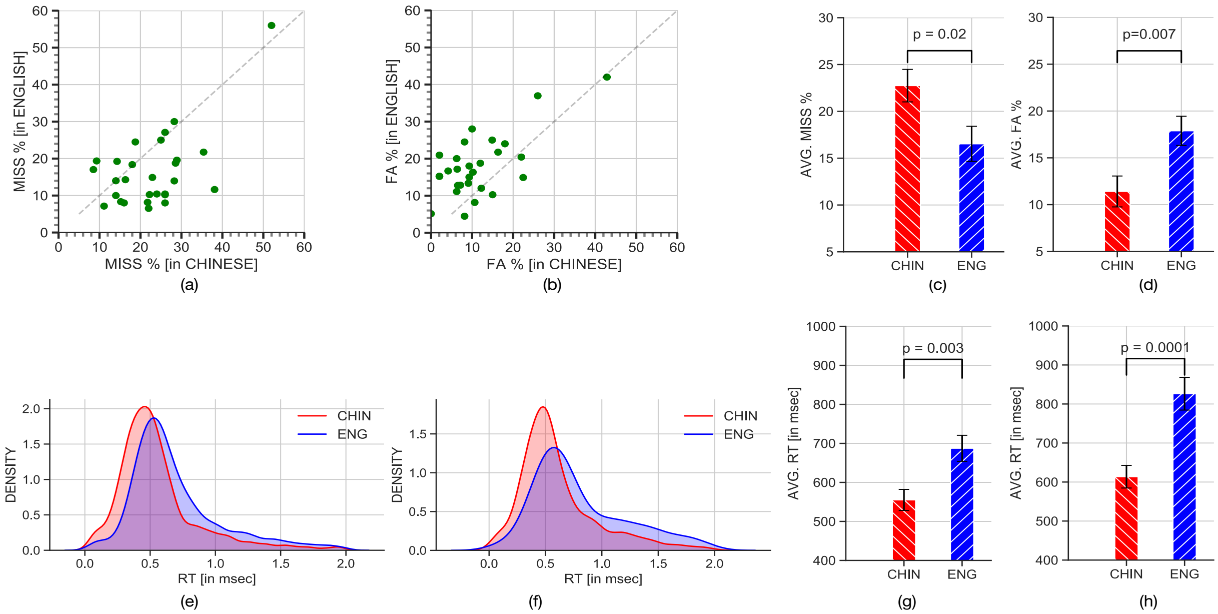


Fig. 2. (color online) Human performance on the talker change detection task, as a function of language familiarity. Panels (a,b): miss and false alarm rates, respectively, for each participant; (c,d): average miss and false rates; (e,f): all participants' pooled response times on correct, and incorrect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

a skewed distribution (see Figure 2(e,f)), the logarthimic transformation of RT was used. The acoustic features included: mel-spectrogram (MEL; using 40 filters), mel-frequency cepstral coefficients (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 msec, with temporal hop of 10 msec) extracted features. 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 Section 3.1), we model RTs seprately 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 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 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 2 talker correct English trials. This is interesting as the x-vector features are designed to capture talker differences and has been shown to be useful in machine diarization systems.

### *3.3 Human-machine comparison*

The machine system performance is shown in Figure 3(b). In a typical evaluation of diarization systems (Ryant *et al.*, 2019), performance is evaluated on long audio recordings with durations up to several minutes. Hence, in the current scenario, where the recordings

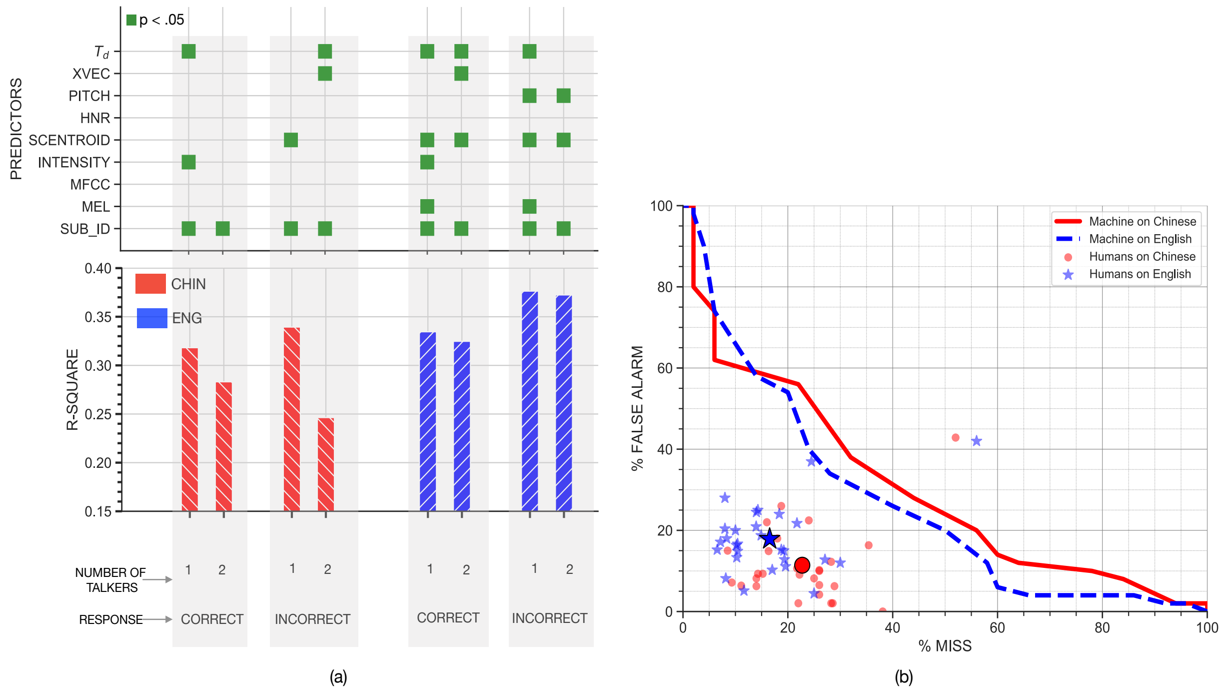


Fig. 3. (color online) (a) Top: Feature significance across models, green square indicates $p < .05$. Bottom: Model $R^2$. (b) Detection error trade-off (DET) curve for the machine system. The scattered points correspond to human participants. The two large shapes correspond to averages across human participants.

range from $5 - 10$ secs in duration, the state-of-the-art diarization system has higher miss and FA rates. Note that the diarization outputs are only analyzed in terms of the number of talkers and not the traditional diarization error rate (DER) metric. Even with this simplified metric, this evaluation shows that, relative to the machine system, the human responses on multi-talker change detection task has (on average) less than half the number of errors. With only a small number of within-talker x-vector embeddings in the stimuli ($6 - 10$ embeddings), the AHC algorithm has considerable difficulty in identfying talker clusters. This performance gap highlights that understanding human processing of talker change detection in short duration recordings can provide important cues for the design of improved talker diarization systems targeting short duration audio signals.

## 4. Discussion

The listening test results show a significant impact of language familiarity on human talker change detection performance. Specifically, the lower miss rate for familiar language suggests 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 langauge trials is significantly higher compared to the unfamiliar langauge trials. This suggests that comprehension of speech (which likely occurs in familiar language stimuli) adversely affects the TCD response time. In the unfamiliar language case, there is no conflict (increased cognitive load) of semantic processing involved.

The regression analysis of RTs indicates 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. The human-machine performance comparison shows that the machine performance is worse than the average performance of the human participants. The traditional diarization system design approach is heavily focused on long duration talker diarization tasks. Our experiments here indicate that more efforts focusing on human understanding of TCD are needed to be directed on diarization of short duration audio signals as these signals are often encountered in conversational settings.

## 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 VK and SG was supported by grants from the British Telecom India Research Center (BTIRC).

## 5. References

## References and links

Boersma, P., and Weenink, D. “Praat: A system for doing phonetics by computer, 2000,” Software available at www. praat. org **4**(2).

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," Cognitive Psychology **98**, 73 – 101.

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

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

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," Psychological Science **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," The Journal of the Acoustical Society of America **127**(4), 2498–2508.

Krishnamohan, V. (**2020 (accessed April 24, 2020)**). *Talker Change Detection Task Demonstration*, www.github.com/iiscleap/langtcd_demo.

Laver, J. D. M. (**1968**). "Voice quality and indexical information," British Journal of Disorders of Communication **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 Communication **30**(1), 9 – 26.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Perception & Psychophysics **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 *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (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, doi: 10.1126/science.1207327.

Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (**2019**). "Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices," The Journal of the Acoustical Society of America **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://coml.lscp.ens.fr/dihard/2018/docs/first_dihard_eval_plan_v1.3.pdf.

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," Proc. of Interspeech 978–982.

Sell, G., Suied, C., Elhilali, M., and Shamma, S. (**2015**). "Perceptual susceptibility to acoustic manipulations in speaker discrimination," The Journal of the Acoustical Society of America **137**(2), 911–922.

Sharma, N., Krishnamohan, V., Ganapathy, S., Gangopadhayay, A., and Fink, L. (**2020**). "On the impact of language familiarity in talker change detection," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6249–6253.

Sharma, N. K., Ganesh, S., Ganapathy, S., and Holt, L. L. (**2019**). "Talker change detection: A comparison of human and machine performance," The Journal of the Acoustical Society of America **145**(1), 131–142.

Singh, P., Vardhan, H., Ganapathy, S., and Kanagasundaram, A. (**2019**). "LEAP diarization system for the second DIHARD challenge," in *Proc. 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 *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process. (ICASSP)*, IEEE, pp. 5329–5333.