ANALYZING HUMAN REACTION TIME FOR TALKER CHANGE DETECTION

Neeraj Sharma[*+], *Shobhana Ganesh*[+], *Sriram Ganapathy*[+], *Lori L. Holt*[*]

[*]Carnegie Mellon University, Pittsburgh, USA, [+] Indian Institute of Science, Bangalore, India

ABSTRACT

The ability to detect a change in the input is an essential aspect of perception. In speech communication, we use this ability to identify "talker changes" when listening to conversational speech (such as, audio podcasts). In this paper, we propose to improve our understanding about how fast listeners detect a change in talker, and the acoustic features tracked to identify a voice by designing a novel experimental paradigm. A listening experiment is designed in which listeners indicate the moment of perceived talker change in multi-talker speech utterances. We examine talker change detection performance by probing the human reaction time (RT). A random forest regression is used to model the relationship between RTs and acoustic features. The findings suggest that: (i) RT is less than a second, (ii) RT can be predicted from the difference in acoustic features of segment before and after change, and (iii) there a exists a significant dependence of RT on MFCC-D1 (delta MFCCs) features between segments of speech before and after the change instant. Further, a comparison with a machine system designed for the same task of TCD using speaker diarization principles showed a poor performance relative to the humans.

Index Terms— Reaction time, talker change detection, speech analysis, random forest regression, speaker diarization.

1. INTRODUCTION

Everyday visual and auditory perception requires that we respond to changes in the incoming input, such as in recognizing a change in the brightness of a screen or a change in lead instrument in a song. Several studies in vision [1, 2, 3] and audition [4, 5, 6, 7, 8] have used parameterized stimuli to demonstrate the strong neurobiological signals elicted in response to a change in sensory input. Change detection becomes especially interesting in the context of everyday speech communication, which involves more than extracting a linguistic message [9]. Listeners also track paralinguistic indexical information in speech signals, such as talker identity, dialect, and emotional state [10]. Indeed, in natural speech communication, linguistic and indexical information are likely to interact since conversations typically involve multiple talkers who take turns of arbitrary duration, with gaps on the order of only 200 ms [11]. On the listener's side, the perception of conversational speech demands quick adjustment to talker changes. This is beneficial for speech processing as perceptual learning of talker identity benefits speech intelligibility in both quiet [12] and in acoustically-cluttered environments [13, 14].

Despite the arguably important role of talker change detection (TCD) in speech communication, there remain many important unanswered questions: how quickly can humans detect talker change? what is the human accuracy in a TCD task? can the reaction time (RT) for TCD be modeled with simple temporal/spectral features?, and what are the acoustic features impacting the RT in TCD?. Gaining understanding of these can benefit design of machine systems for conversational speech recognition. To address these questions, this paper presents analysis of data collected from a listening test based study on TCD. The experiment setup for the listening task is described in Section 2 and the human performance is analyzed in Section 3. We attempt modeling RT prediction from acoustic features in Section 4. In Section 5, we use a machine system for the same task. The paper concludes in Section 6. Recently, in [15], we have presented the full details of the TCD experiment without much focus on predicting RT. This work expands the analysis of human data and prediction of psychophysical data using random forest regression.

The work done was supported partly by the BrainHub research grant from CMU and from Pratiksha Trust Young Investigator Award.

2. EXPERIMENT SETUP

Fig 1 provides an illustration of the setup. Each stimulus was composed of two concatenated utterances sourced from audio books featuring natural speech intonations, and spoken by either a single male talker or two different male talkers (denoting by T_x and T_y). The utterances were always drawn from different stories, or parts of a story, so that semantic continuity did not provide a clue to talker continuity, and were taken from five audio books drawn from the LibriSpeech corpus [16], a public-domain corpus of audio data corresponding to audio books read by different talkers. Based on an informal pilot experiment aimed at finding a set of perceptually separable voices, we chose five talkers from the corpus (IDs 374, 2843, 5456, 7447, 7505) for the listening test stimulus design (here referred to as T1, T2, etc.). To make a stimulus, talker T_x was chosen from the list of N talkers, and a sentence utterance was retrieved from the corresponding talker's audio book. A short utterance from another talker T_y was chosen, and this was concatenated to the utterance from T_x. As the utterances were natural speech, there were natural pauses. Owing to this, the silent interval between T_x's end and T_y's start after concatenation was random and ranged from $200 - 1000$ ms. In any stimulus, speech corresponding to T_x was between $5 - 10$ s and that corresponding to T_y was 4 s. A sample stimulus is shown in Fig. 1. For each pair of T_x-T_y talkers, there were $M = 8$ unique stimuli. This resulted in a total of $M \times N^2 = 200$ distinct speech stimuli, each $9 - 14$ s in duration.

The listeners responded with a button press upon detecting a talker change, thus providing a continuous reaction time measure of how much of an acoustic sample was needed to detect a change in talker. To the best of our knowledge, this is the first application of a RT change-detection approach to examine human TCD performance. The study was carried out in isolated sound booths and over headphone listening, and using a GUI developed with Gorilla, a software platform for designing behavioral science tests.

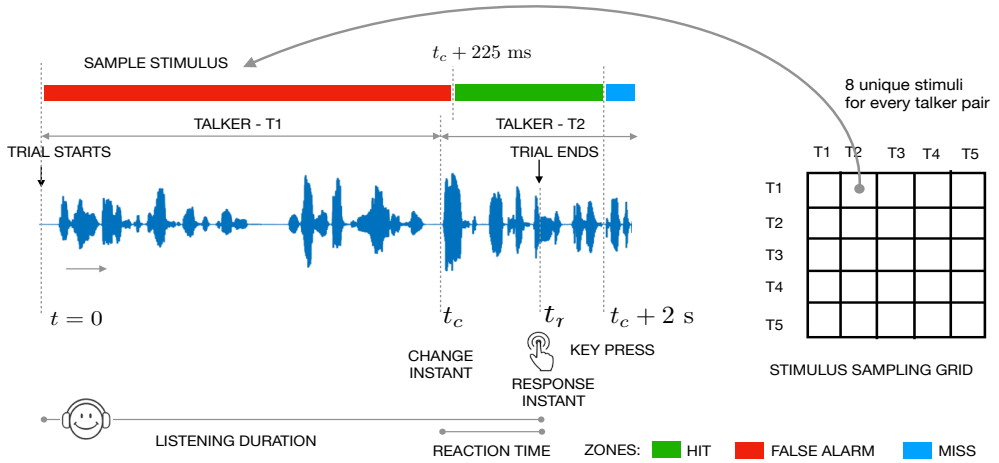


Fig. 1. Illustration of the proposed talker change detection (TCD) paradigm used in the present listening test study.

3. HUMAN PERFORMANCE IN TCD

A total of 17 subjects, self reported as normal hearing and conversant with English, took part in the study. A subject on average took 45 mins for the complete task. For each trial the RT for change detection was obtained as the difference between the response instant (denoted by t_r, instant of button press) and the ground-truth acoustic change instant (denoted by t_c), that is, $RT = t_r - t_c$. The lower limit for RT for change perception in sound attributes is of the order of $RT < 250$ ms [17]. Hence, RTs in the range $0 - 250$ ms are likely to be associated with speech heard prior to the change instant t_c. The upper bound on RT (2000 ms) was chosen based on prior research [7].

The 200 trials per subject were categorized into two pools for analyses: *Pool A* - involving trials with T_x a different talker from T_y (two-talker trials) and either $RT > 225$ ms or no button press, and *Pool B* involving trials with $T_x = T_y$ and trials with $T_x \neq T_y$ but $RT < 225$ ms. These all are single-talker trials (i.e the trials in which the subject's response was based on attention to only one talker). From these pools of data, we defined the following detection measures:

- **Hit rate:** A hit corresponds to a trial in *Pool A* with 225 ms $< RT <$ 2000 ms. Hit rate is the ratio of number of hits to the number of trials in *Pool A*.
- **Miss rate:** A miss corresponds to a trial in *Pool A* with $RT >$ 2000 ms. Miss rate is the ratio of number of misses to the number of trials in *Pool A*. Note that the miss rate is 100 - hit rate.
- **False alarm rate:** A false alarm (FA) corresponds to a trial in *Pool B* featuring a button press. False alarm rate is the ratio of number of FAs to the sum of trials in *Pool B* and *Pool A* (this equals 200).

Fig 2(a) depicts the distribution of TCD reaction time t_r as a function of ground-truth talker change instant t_c for all trials which have a talker change (taken from *Pool A* and *Pool B*). As seen, the majority (approx. 95%) of responses fall in the hit zone, that is, $t_c + 225 < t_r < t_c + 2000$ ms. Analyzing the hit trials from *Pool A*, the subject-wise RT summary is shown in Fig 2(b). Across subjects, the response time to detect a talker change tended to require mostly under a second of speech from the true change instant, with subject-dependent distributions of average RT and variability across quantiles. Analyzing the detection parameters, the subject-wise hit, miss and FA rates are shown in Fig 2(c). The hit, miss, and false alarm rates averaged across all subjects were 97.38%, 2.62%, and 8.32%, respectively. The listeners performed the TCD task very accurately; the average d-prime[1] across subjects was 3.48.

4. MODELING RT FOR TCD

We explored the dependence of RT on the acoustic features in the speech segments before and after change instant. This is illustrated in Fig 3, and the segments are denoted by D_b (before change instant) and D_a (after change instant), respectively. We considered a set of acoustic features depicted in Table 1. These features are computed every 10 ms with *Hanning* windowed short-time segments of 25 ms. All features were extracted using the *Yaafe* [18] Python package, an efficient open-source code library for speech and audio analysis.

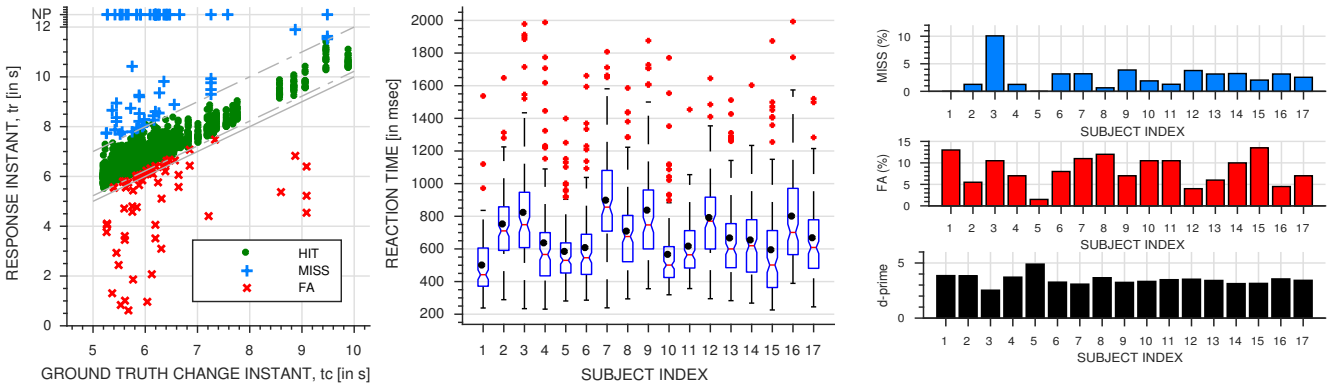
For each feature set, we summarized the segments D_b and D_a using the mean of the features in each segment. The PLOUD, and SPECT feature set were characterized by a combination of different features. Hence, we mean- and variance-normalized these feature sets over the whole duration prior to segment-wise mean computation. Following this, we computed the Euclidean distance between the obtained means. Owing to significant variability in RT across subjects (see Fig 2(b)), we modeled each subject's RT separately.

To model the dependence of RT on acoustic features we used Random Forest based regression [19]. This approach is devoid of any assumption on linear relationship between the dependent (RT) and independent variables (the acoustic feature set), and hence, suits the exploratory analysis. We used the *Scikit-learn* Python package for implementation. The best performance was obtained with number of estimators set to 40, and minimum samples per leaf set to 5 (allowing reduced overfitting). All implementation used $k-$fold validation to evaluate the performance. The performance metric was the "% explained variance", a value close to 100% indicating perfect prediction.

4.1. Results

Fig 4(a) depicts the subject-wise performance. The validation set performance (with $k = 10$) is subject dependent. For some subjects

[1] d-prime is defined as $\mathcal{Z}$(hit rate) $-$ $\mathcal{Z}$(FA rate), where function $\mathcal{Z}$(p), p $\in [0, 1]$, is the inverse of the cumulative distribution function of the Gaussian distribution



(a) (b) (c)

Fig. 2. (a) Illustration of human reaction time (RT) versus the ground-truth talker change instant (t_r vs t_c) across a total of 2720 trials (with $T_x \neq T_y$) over 17 subjects. The three inclined gray lines from bottom to top correspond to $t_r = t_c$, $t_c + 225$, $t_c + 2000$, respectively. NP stands for no button press. (b) Subject-wise summary using a boxplot of RTs in trials with hits. The black dots correspond to means. (c) Subject-wise miss and false alarm rates, and d-prime obtained from 200 trials for each subject.

Table 1. Acoustic features used in modeling RT.

FEATURE SET	FEATURES	TYPE	DIMENSION	TIME SCALE
F0	Fundamental Frequency	Spectral	1×1	25 ms
LSF	Line spectral frequencies	Spectral	10×1	25 ms
MEL	Mel-spectrogram	Spectral	40×1	25 ms
MFCC	Mel-frequency cepstral coefficients	Spectral	12×1	25 ms
MFCC-D1	First-order temporal derivative of MFCCs	Spectral	12×1	25 ms
MFCC-D2	Second-order temporal derivative of MFCCs	Spectral	12×1	25 ms
TEMP	Derivative of short-time energy	Temporal	1×1	25 ms
PLOUD	Loudness strength, sharpness, and spread	Spectral	3×1	25 ms
SPECT	Spectral flatness, Spectral flux, Spectral roll-off, Spectral shape, Spectral slope	Spectral	8×1	25 ms

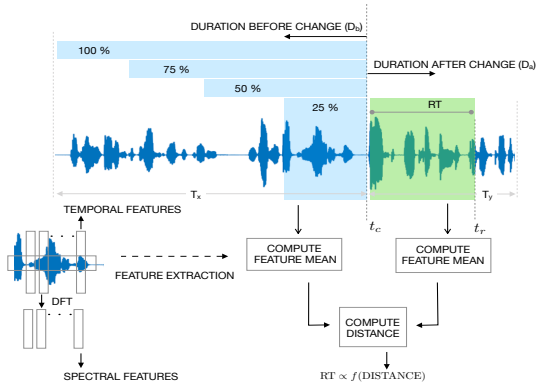


Fig. 3. Approach to modeling RT using acoustic features before and after the talker change instant.

the score is more than 40% (for instance, 65% for subject 15). Pooling data from all subjects, the prediction score was close to 55%. Interestingly, we also found a relative improvement in prediction score with increase in segment duration before the change instant (D_b). Pursuing a feature importance analysis (based on Gini importance [19]), we found that MFCC-D1 feature maximally impacted the decisions in the random forest. We analysed this further using a least square fit to the RT data, separately on each feature type seprately. The slope was found to be maximum (and of negative sign) for RT

Table 2. Human versus machine diarization performance.

System	**Miss-rate** (%)	**FA rate** (%)
Human	2.62	8.35
Machine Thresh. 1 [20]	2.62	11.35
Machine Thresh. 2 [20]	10.8	8.32

versus MFCC-D1. This is shown in Fig 5, and hints at lower RT for stimuli with higher MFCC-D1 distance between D_a and D_b segments. An illustration of the predicted RTs obtained using the random forest regression is shown in Fig 6. The prediction SNR in both training and validation set was found to be close to 13 dB. The low SNR is not surprising considering the high variability in the RTs. However, we got a decent score on %explained variance (shown in Fig 4) and the Spearman's rank correlation was found to be around 0.76 on the validation set instances.

5. HUMAN VERSUS MACHINE DIARIZATION

We evaluated the performance of an offline diarization with the same stimulus materials used in the human TCD experiment. The system was designed to segment audio into distinct talker segments based on i-vector and probabilistic linear discriminant analysis (PLDA) [20]. Subsequently, the talker change instants can be obtained as segment boundaries. The complete system setup was developed using the Kaldi toolkit [21]. This involved training the i-vector ex-

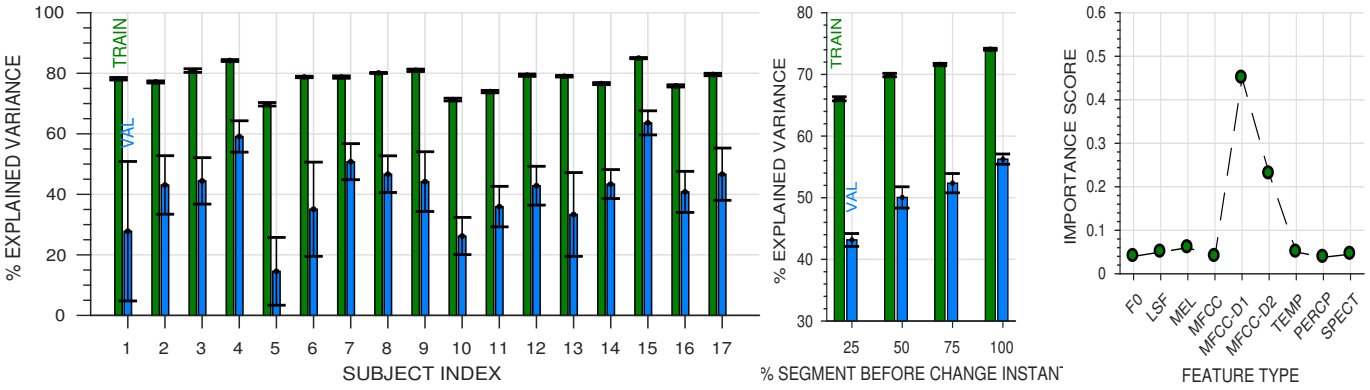
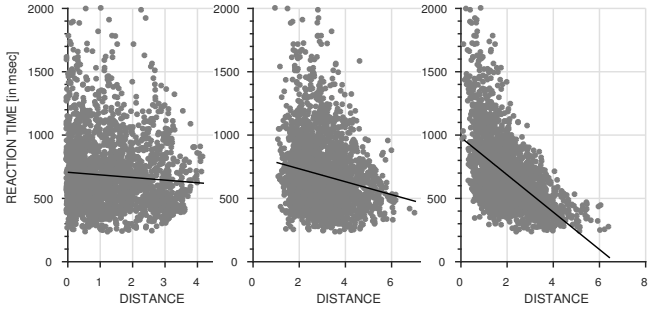
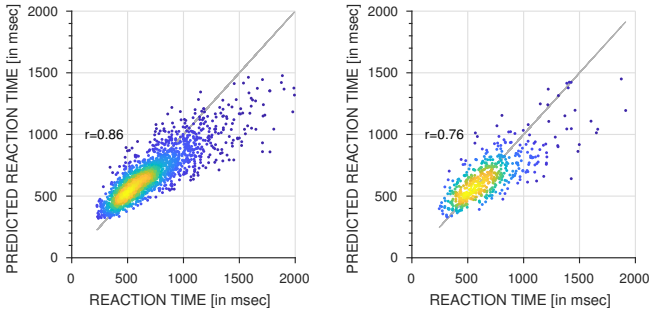


Fig. 4. Illustration of reaction time modeling performance obtained with Random Forest regression. (a) Subjectwise training and validation performance, obtained by 10 fold cross-validation. (b) Train and validation performance on pooled data from all subjects, obtained by 5 fold cross-validation. (c) Feature importance score, obtained using node impurity metric.



(a) F0 (b) MFCC (c) MFCC-D1

Fig. 5. Illustration of RT as a function of feature distance. The line in the plots depicts a least sqaure fit to the data.



(a) On training set. (b) On validation set.

Fig. 6. Illustration of predicted RTs obtained with random forest regression. The color indicates density of samples points, yellow to blue indicating more to less.

tractor based on a universal background model composed of a 512-component Gaussian mixture model with a diagonal covariance matrix and trained on the LibriSpeech corpus [16]. The system used 12-dimensional MFCC features, obtained from successive 25 ms (with temporal shifts of 10 ms) short-time segments derived from the audio files. The MFCC features were mean- and variance-normalized using a 3 s running window. The i-vector representations were 128-dimensional. The pairwise PLDA scores computed between 1 s i-vector segments are clustered using agglomerative hierarchical clustering (AHC) [20]. The audio files corresponding to talkers used in the listening test were removed from the training dataset.

The machine system can be operated at any point in the detection-error-tradeoff contour while the human results are found at one operating point (average results from all the subjects). In order to make the direct comparison between the human and machine systems, we found the threshold of the machine system that matched the miss rate of the human system (Thres. 1) and compared the false-alarm rate at this operating point. Similarly, the threshold obtained by matching the false-alarm-rate (Thres. 2) allows the comparison of the miss-rate between human and machine systems. This comparison is reported in Table 2. Comparing human and machine systems, we find a considerable gap in performance, with humans significantly outperforming a state-of-the-art machine system.

6. DISCUSSION AND CONCLUSION

The findings suggest an average RT of 680 ms (with std. dev. 270 ms) for TCD in the designed task. The human accuracy was found to be good with low false alarm and miss detection across subjects. The possibility of RT modeling was evaluated and score close to 55% was obtained with a simple random forest based regression approach. Further, we found that the rate of change of MFCCs impacted the regression more than other features. This is interesting as MFCC are associated with capturing the formant information, and the rate of change further captures the modulation in formant frequencies. Interestingly, the prediction performance improved with segment duration before change instant hinting at long-term statistics being modeled by the listeners in this task. Application of a state-of-the-art machine system on the same task showed a significant performance gap when compared to humans.

Past studies have used RT to analyze perception of simpler acoustic attributes. For example, studies of tone onset detection[4] and broadband sound onset [6, 17] have reported an inverse relationship between RT and stimulus loudness / spectral bandwidth. The presented findings are first in the direction of using RT analysis approach to study talker change detection. As future work, we consider that an EEG study capturing the response of brain to talker change instants on a similar task can give further insights into online talker modeling while listening. On the machine system design, use of acoustic features correlating with RT predictions may benefit diarization of conversational speech recordings. Stimulus samples and analysis codes are hosted at [22].

7. REFERENCES

[1] Elias H. Cohen, Elan Barenholtz, Manish Singh, and Jacob Feldman, "What change detection tells us about the visual representation of shape," *Journal of Vision*, vol. 5, no. 4, pp. 3, 2005.

[2] Amin Mirzaei, Seyed-Mahdi Khaligh-Razavi, Masoud Ghodrati, Sajjad Zabbah, and Reza Ebrahimpour, "Predicting the human reaction time based on natural image statistics in a rapid categorization task," *Vision Research*, vol. 81, pp. 36 – 44, 2013.

[3] R. T. Pramod and S. P. Arun, "Do computational models differ systematically from human object perception?," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Larry E. Humes and Jayne B. Ahlstrom, "Relation between reaction time and loudness," *Journal of Speech, Language, and Hearing Research*, vol. 27, no. 2, pp. 306–310, 1984.

[5] Clara Suied, Patrick Susini, Stephen McAdams, and Roy D. Patterson, "Why are natural sounds detected faster than pips?," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. EL105–EL110, 2010.

[6] Josef Schlittenlacher, Wolfgang Ellermeier, and Gül Avci, "Simple reaction time for broadband sounds compared to pure tones," *Attention, Perception, & Psychophysics*, vol. 79, no. 2, pp. 628–636, Feb 2017.

[7] Yves Boubenec, Jennifer Lawlor, Urszula Górska, Shihab Shamma, and Bernhard Englitz, "Detecting changes in dynamic and complex acoustic environments," *ELife*, vol. 6, 2017.

[8] Golbarg Mehraei, Barbara Shinn-Cunningham, and Torsten Dau, "Influence of talker discontinuity on cortical dynamics of auditory spatial attention," *NeuroImage*, vol. 179, pp. 548 – 556, 2018.

[9] Stephen D. Goldinger, "Echoes of echoes? An episodic theory of lexical access," *Psychological Review*, vol. 105, no. 2, pp. 251–279, 1998.

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

[11] Stephen C. Levinson, "Turn-taking in human communication - Origins and implications for language processing," *Trends in Cognitive Sciences*, vol. 20, no. 1, pp. 6 – 14, 2016.

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

[13] Pdraig 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.

[14] Ingrid S. Johnsrude, Allison Mackey, Hlne 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.

[15] 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*, vol. 145, no. 1, pp. 131–142, 2019.

[16] 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.*, April 2015, pp. 5206–5210.

[17] David S. Emmerich, Deborah A. Fantini, and Wolfgang Ellermeier, "An investigation of the facilitation of simple auditory reaction time by predictable background stimuli," *Perception & Psychophysics*, vol. 45, no. 1, pp. 66–70, Jan 1989.

[18] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard, "Yaafe, an easy to use and efficient audio feature extraction software.," in *ISMIR*, 2010, pp. 441–446.

[19] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 413–417.

[21] 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," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2011, number EPFL-CONF-192584.

[22] Neeraj Kumar Sharma, Shobhana Ganesh, Sriram Ganapathy, and Lori L. Holt, "Resources used in the talker change detection study," `https://github.com/neerajww/icassp2019_rt_prediction`, (Feb. 2019).