Talker change detection: A comparison of human and machine performance

Neeraj Kumar Sharma, Shobhana Ganesh, Sriram Ganapathy, and Lori L. Holt

Citation: The Journal of the Acoustical Society of America **145**, 131 (2019); doi: 10.1121/1.5084044
View online: https://doi.org/10.1121/1.5084044
View Table of Contents: https://asa.scitation.org/toc/jas/145/1
Published by the Acoustical Society of America

Talker change detection: A comparison of human and machine performance

Neeraj Kumar Sharma,[1,a)] Shobhana Ganesh,[2] Sriram Ganapathy,[2] and Lori L. Holt[1]
[1]*Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA*
[2]*Department of Electrical Engineering, CV Raman Road, Indian Institute of Science, Bangalore 560012, India*

(Received 15 August 2018; revised 25 October 2018; accepted 1 December 2018; published online 7 January 2019)

The automatic analysis of conversational audio remains difficult, in part, due to the presence of multiple talkers speaking in turns, often with significant intonation variations and overlapping speech. The majority of prior work on psychoacoustic speech analysis and system design has focused on single-talker speech or multi-talker speech with overlapping talkers (for example, the cocktail party effect). There has been much less focus on how listeners detect a change in talker or in probing the acoustic features significant in characterizing a talker's voice in conversational speech. This study examines human talker change detection (TCD) in multi-party speech utterances using a behavioral paradigm in which listeners indicate the moment of perceived talker change. Human reaction times in this task can be well-estimated by a model of the acoustic feature distance among speech segments before and after a change in talker, with estimation improving for models incorporating longer durations of speech prior to a talker change. Further, human performance is superior to several online and offline state-of-the-art machine TCD systems.
 https://doi.org/10.1121/1.5084044

[JFL] Pages: 131–142

I. INTRODUCTION

Everyday speech communication involves more than extracting a linguistic message.[1] Listeners also track paralinguistic indexical information in speech signals, such as talker identity, dialect, and emotional state.[2] 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.[3] On the listener's side, the perception of conversational speech demands quick perception of talker changes to support communication.

Perceptual learning of talker identity enhances speech intelligibility in both quiet[4] and acoustically cluttered environments.[5,6] This suggests that sensitivity to talker attributes affects speech recognition both in clear speech and under adverse listening conditions. Further, talker dependent adaptability in perception can be induced from exposure to just a few sentences.[7] 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 would seem to rely upon an ability to track regularities in the perceived features specific to a voice, and to detect changes from these upon a talker change. Lavner *et al.* (2009)[8] suggest that the talkers are identified by a distinct group of acoustic features. Yet, Sell *et al.* (2015)[9] argue that a combination of vocal source, vocal tract, and spectro-temporal receptive field[10] features fail to explain perceived talker discrimination in a listening test with simple single-word utterances. In a similar way, Fenn *et al.*[11] have described inattention to talker changes in the context of listening for comprehension as a form of talker change deafness.[12] They suggest that voice information is not continuously monitored at a fine-grain level of acoustic representation, and conversational expectations may shape the way listeners direct attention to voice characteristics and perceive differences in voice. In fact, Neuhoff *et al.* (2014)[13] found improved voice change detection when the language is unfamiliar to the listener, suggesting that there may be interactions between linguistic and indexical information.

Acknowledging that conversational demands[14] in natural speech will often shift attention toward acoustic features that signal linguistic rather than indexical information, listeners' ability to detect talker changes does suggest that they track the variability in acoustics features associated with a talker's voice. Yet, despite the importance of indexical characteristics of speech to communication, quite little is known about the nature of the detailed acoustic features across which talkers differ, the distributions of information characterizing different talkers along these acoustic features[15,16] and the listeners' ability to detect a change in talker. This is especially true for fluent, connected speech, as opposed to isolated words. In this paper, we aim to advance understanding of the information human listeners use to track the change in talker in continuous multi-party speech. We first develop and test a novel experimental paradigm to examine human talker change detection (TCD) performance. We next model the listeners' reaction time (RT) to respond to a talker change in relationship to multiple acoustic features as a means of characterizing the acoustic feature space that listeners may track in a voice. We then relate these human

[a)]Electronic mail: nsharma2@andrew.cmu.edu

perceptual results with performance of state-of-the-art online and offline machine systems implementing TCD.

A. RT as a measure of TCD

We developed a novel paradigm with the goal of obtaining a continuous behavioral measure of listeners' ability to detect a change in talker across relatively fluent, continuous speech. Each stimulus was composed of two concatenated 9–14 s utterances sourced from audio books, and spoken by either a single male talker or two different male talkers. 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. Listeners responded with a button press upon detecting a talker change, thus providing a continuous RT measure of how much of an acoustic sample was needed to detect a change in talker. Figure 1 provides an illustration of the paradigm. To the best of our knowledge, this is the first application of a RT change-detection approach to examine human TCD performance.

Past studies have used RT to analyze perception of simpler acoustic attributes. For example, studies of tone onset detection[17] and broadband sound onset[18,19] have reported an inverse relationship between RT and stimulus loudness/spectral bandwidth. Studies guiding the design of warning sounds[20,21] have shown faster detection of natural, compared to synthetic stimuli. Particularly relevant to the present study, recent research characterizes human listeners' ability to track distributional noise statistics in an audio stream by asking listeners to report a change in statistics with a button press.[22] Detection of a change was facilitated when listeners heard longer noise samples. This prior work illustrates the promise of the RT to detect a change as a means by which to examine how listeners build a model of incoming sound regularities, although speech signals are inherently more non-stationary, and the distributional acoustic information that contributes to talker identity is not well understood.

B. Comparing machine and human performance in TCD

With an increasing repository of conversational audio data,[23,24] automatic detection of talker change is considered to be an essential preprocessing in machine speech recognition.[25] For example, the availability of time stamps corresponding to talker change instants in a recording benefits both speaker identification[26] and speech recognition[27] tasks. Recent results from the machine systems literature on TCD[28–32] suggest that TCD is difficult. The state-of-the-art in TCD can be categorized into two approaches: metric-based and classification-based. The metric-based approach relies on computing a distance, in some feature space, between successive short-time segments (such as 25 ms segments, with successive shifts 10 ms) of a speech recording. A talker change is flagged upon detection of a distance that exceeds a preset threshold. The accuracy of TCD in metric-based approaches is dependent on the choice of features (such as vocal tract features[33,34] and vocal source features[30,35–37]), segment duration (usually >3–5 s), and the distance metric [such as likelihood ratio,[38] Bayesian information criterion (BIC),[33,39] or improved BIC[40]]. An alternative is the classification-based approach whereby a binary classifier is trained to classify short-time speech segments as talker change or no talker change. The accuracy of the approach is dependent on the feature space, the complexity of the classifier (such as support vector machines,[41] neural networks,[30] and deep neural networks[31,32]), and the amount of training data. In the present work, we evaluate the performance of a few state-of-the-art machine systems across the same stimuli and with the same performance metrics as used in the human perceptual paradigm. These comparisons may help in identifying whether there are performance gaps in human and machine TCD across fairly continuous, fluent speech.

II. METHODS

A. Participants

A total of 17 participants in the age group of 18–36 yr (university students and one staff member, median age 26 yr) took part in the listening test. All listeners reported normal hearing with good fluency in speaking and understanding English. All participants provided informed consent to participate in the study, and the study protocol was approved by the Carnegie Mellon University Institutional Review Board and the Indian Institute of Science Review Board.

B. Stimuli

The stimuli were composed of concatenated utterances from two talkers, talker T_x and talker T_y. The utterances were

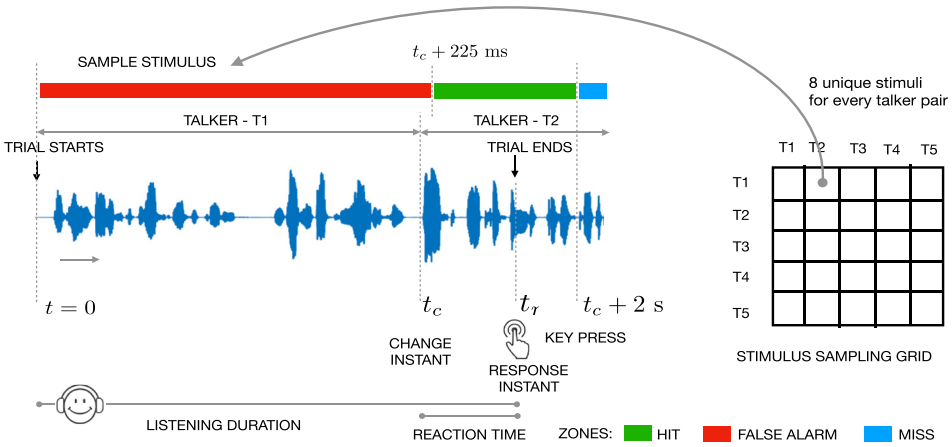

FIG. 1. (Color online) Illustration of the proposed TCD paradigm used in the present listening test study.

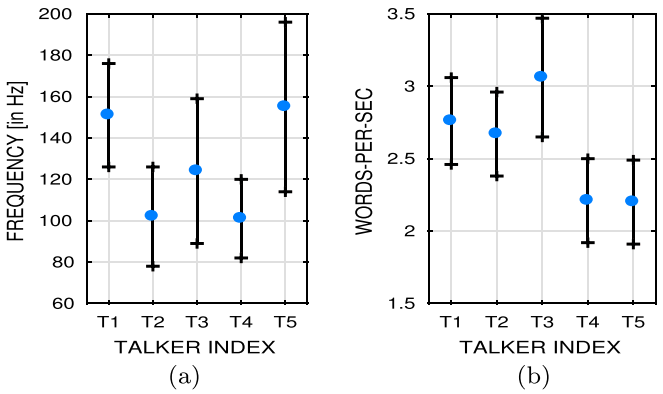

FIG. 2. (Color online) A comparison of talker attributes with respect to (a) fundamental frequency variation and (b) word speaking rate. The vertical bars indicate one standard deviation spread around the mean value.

taken from audio books drawn from the LibriSpeech corpus,[42] a public-domain corpus of about 1000 h of audio data by approximately 1000 distinct talkers. Each trial involved a stimulus that was a concatenation of two utterances drawn from the same, or two different, talkers. The utterances corresponded to sentences read out in the audio book, and featured natural speech intonation and a rise and fall in speech envelope at the start and end. The sentences were chosen randomly from the respective talker's audio book. To avoid an obvious TCD due to gender attributes, we chose all male talkers. Based on an informal pilot experiment aimed at finding a set of perceptually separable voices, we chose five talkers from the corpus [identifications (IDs) 374, 2843, 5456, 7447, and 7505] for the listening test stimulus design (here, referred to as T1, T2, etc.). The average fundamental frequency (estimated using STRAIGHT[43]) and speaking rate expressed as words spoken per second (estimated from the audio book transcripts) of the five talkers are depicted in Fig. 2. This shows significant overlap across talkers, making TCD challenging.

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 to 1000 ms. In any stimulus, speech corresponding to T_x was between 5 and 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. The stimulus set can be represented as sampled from a grid, as shown in Fig. 1. This resulted in a total of $M \times N^2 = 200$ distinct speech stimuli, each 9–14 s in duration.

C. Protocol and apparatus

A graphical user interface (GUI) for stimulus presentation was made using Gorilla,[44] a software platform for designing behavioral science tests. A test session comprised three tasks carried out in sequence, namely, tasks-1, -2, and -3. Task-1 measured RT for noise-to-tone detection and task-2 measured RT for tone frequency change detection (see

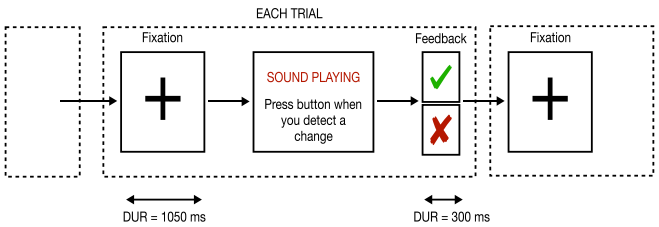

FIG. 3. (Color online) An illustration of a listening test trial.

supplementary material[45]). The acoustic attributes associated with a change in the stimuli in these first two tasks were easily recognizable. As a result, task-1 and task-2 served as benchmarks against which to compare human RT on task-3, associated with TCD.

In each task, listeners were instructed to press a button (space bar) immediately upon detection of a change in the stimulus. The audio stopped after the button press and visual feedback indicating "[✓]" (or "X") for "correct" (or "incorrect") detection appeared immediately (as shown in Fig. 3). Participants were seated in sound-attenuated booths wearing Sennheiser headphones[46] (with flat spectral response from 60 to 8000 Hz) with diotic stimulus presentation. In order to prevent fatigue, participants were allowed to take breaks after blocks of 25 trials. Each participant listened to a few (8–10) different task-3 TCD stimuli (not used in the test) to become conversant with the paradigm. On average, the complete test session was 45 min (with task-1, task-2, and task-3 taking 5, 10 and 30 min, respectively).

D. Performance measures

For each trial in which there was a button press, the RT for change detection was obtained as the difference between the response instant (denoted by t_r) and the ground-truth acoustic change instant (denoted by t_c), that is, $RT = t_r - t_c$. An illustration is provided in Fig. 1. The lower limit for RT for change perception in sound attributes is on the order of $RT < 225$ ms.[19] Hence, RTs in the range 0–225 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.[22]

We analyzed the hits, misses, and false alarms (FAs) in the responses. The 200 trials in task-3 per subject were categorized into 2 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.
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 listening 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*.

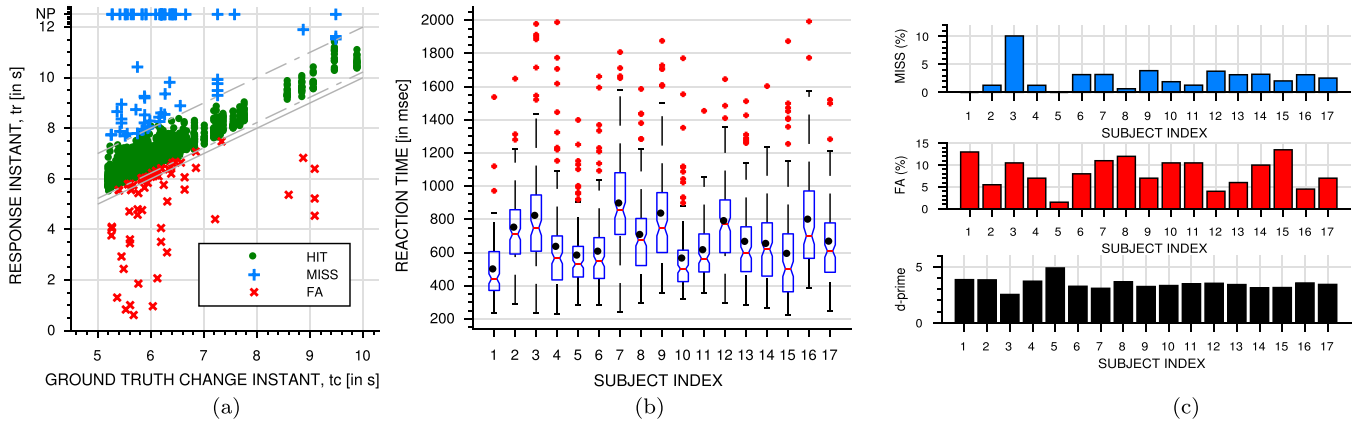

FIG. 4. (Color online) (a) Illustration of human RT versus the ground-truth talker change instant (t_r versus t_c) across a total of 2720 trials (with $T_x \neq T_y$) over 17 subjects. The three inclined gray lines from bottom to top correspond to $t_r = t_c$, $t_c + 225$, $t_c + 2000$, respectively. 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 FA rates, and d' obtained from 200 trials for each subject.

- 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.
- FA rate: A FA corresponds to a trial in *pool B* featuring a button press. FA rate is the ratio of number of FAs to the sum of trials in *pool B* and *pool A* (this equals 200).

III. RESULTS: HUMAN TCD EXPERIMENT

A. RT distribution

Figure 4(a) depicts the distribution of TCD RT t_r as a function of ground-truth talker change instant t_c for all trials that have a talker change (taken from *pool A* and *pool B*). As seen, the majority ($\sim$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. 4(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. 4(c). The hit, miss, and FA rates averaged across all subjects were 97.38%, 2.62%, and 8.32%, respectively. Listeners performed the TCD task very accurately; the average d' across subjects was 3.48 (d' is defined as $\mathcal{Z}(\text{hit rate}) - \mathcal{Z}(\text{FA rate})$, where function $\mathcal{Z}(p)$, $p \in [0, 1]$, is the inverse of the cumulative distribution function of the Gaussian distribution).

The distribution of RT corresponding to hits from all subjects is shown in Fig. 5. The non-Gaussian nature of the data is evident from the histogram and the normal probability plot. To improve the Gaussianity, we applied a log transformation ($\log_{10}$RT) on the RT data; the resulting distribution is shown in Fig. 5. We used this transformed data in the regression analysis, which is presented next.

B. Dependence of RT on speech duration

We examined the extent to which the duration of speech experienced prior to a talker change impacted TCD. We probed this using linear regression analysis on log RT (averaged across subjects) versus speech duration before change instant t_c. As the stimulus set is composed of naturally spoken sentences from story books, the grid along speech duration is non-uniformly sampled in the dataset. Hence, we performed the regression analysis only for stimuli with $t_c < 7000$ ms, as beyond this t_c value, the grid was sparsely sampled. The result is shown in Fig. 6. The high variability in logRT may be attributed to the complicated feature space associated with speech signals. This is unlike what is observed in tone frequency (or noise-to-tone) change detection for which the variability in RT correlates with the simple spectral difference (see supplementary material[45] and also Ref. 22). Despite considerable variability, we observe a trend depicting TCD is slower with shorter samples of speech from the initial talker (a decrease in the t_c value). Next, we attempt to model RT as a function of different acoustic features extracted from the stimulus before and after change instant to understand the acoustic dimensions listeners may track in TCD.

C. Modeling RT with acoustic features

Past studies[47,48] have used RT modeling to reveal visual input features associated with object recognition. Motivated by this, here, we model RT to detect a change in talker across hit trials as a function of the difference in the acoustic feature space between speech segments sampled before and after the change instant. We consider a collection of acoustic features that may be important in tracking voice identity, and examine their ability to estimate human participants' TCD RT.

The approach is illustrated in Figs. 7(a) and 7(b). Let D_b and D_a denote segments of the stimulus before and after change instant t_c, respectively. We hypothesize that a listener estimates acoustic features from these segments, summarizes them, and compares the summary using a distance measure. The resulting distance serves as strength of evidence for change detection, and by Piéron's law[49] this should impact the RT. The greater the distance, the faster listeners may detect a talker change, and thus the smaller the RT value.

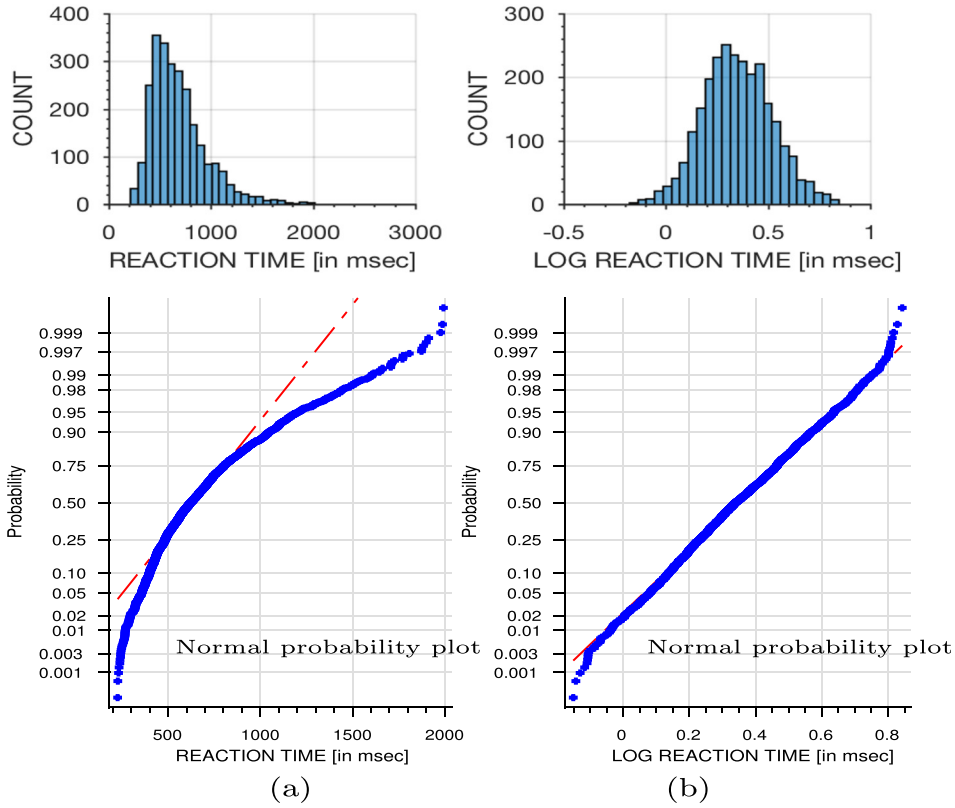

FIG. 5. (Color online) Illustration of the distribution (obtained as a histogram) of the RT data for trials on which there was a hit for (a) raw RT data and (b) log-transformed RT data to improve the fit to a normal distribution.

We assume D_a corresponds to the segment from t_c to t_r. For D_b, we use different segment durations before the change instant (that is, 25%, 50%, 75%, and 100% of t_c duration segment before change instant t_c).

1. Feature computation

A wide range of acoustic features can be computed from D_b and D_a segments. Here, we consider the nine sets of features described in Table I. These are chosen to sample from a range of possible temporal and spectral acoustic features [illustrated in Fig. 7(c)]. From the perspective of speaker attributes, the feature set can be further grouped into those capturing pitch ($F0$), vocal tract formants (line spectral frequencies, LSFs; mel-spectrogram, MEL; mel-frequency cepstral coefficients, MFCCs), rate of temporal variation of vocal tract features [temporal derivatives[50] of MFCCs, namely, first-order temporal derivative of MFCCs (MFCC-D) and second-order temporal derivative of MFCCs (MFCC-DD)], perceived loudness (PLOUD[51]), spectral timbre (SPECT), and the rate of temporal variation in short-time energy (ENGY-D). These features are computed every 10 ms with *Hanning* windowed short-time segments of 25 ms. All features were extracted using the *Yaafe*[52] Python package, an efficient open-source code library for speech and audio analysis.

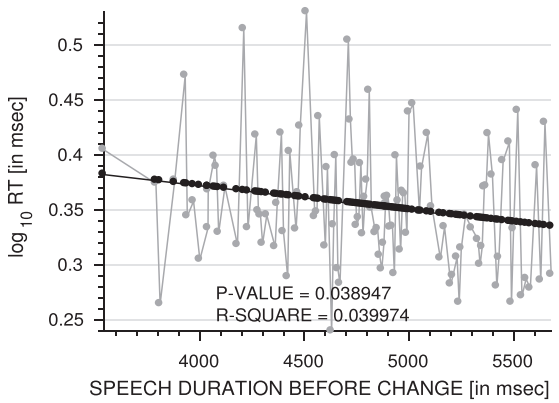

FIG. 6. Dependence of average RT on speech duration before the change instant. The black line is the linear regression fit.

2. Regression on feature distance

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 sets 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. 4(b)], we modeled each subject's RT separately.

We defined the number of trials with a hit for the pth subject to be denoted by N_p. Corresponding to these trials, we have N_p RTs, and for each RT we compute a distance between the mean features extracted from segments D_b and D_a. There are nine such distances based on the choice of features, and we denote these by d_k, $k = 1,\ldots,9$, that is, one distance for each feature set. To evaluate the impact of the feature distances on RT, we perform a linear regression on feature distances to estimate the RT using a regression model for the kth feature set

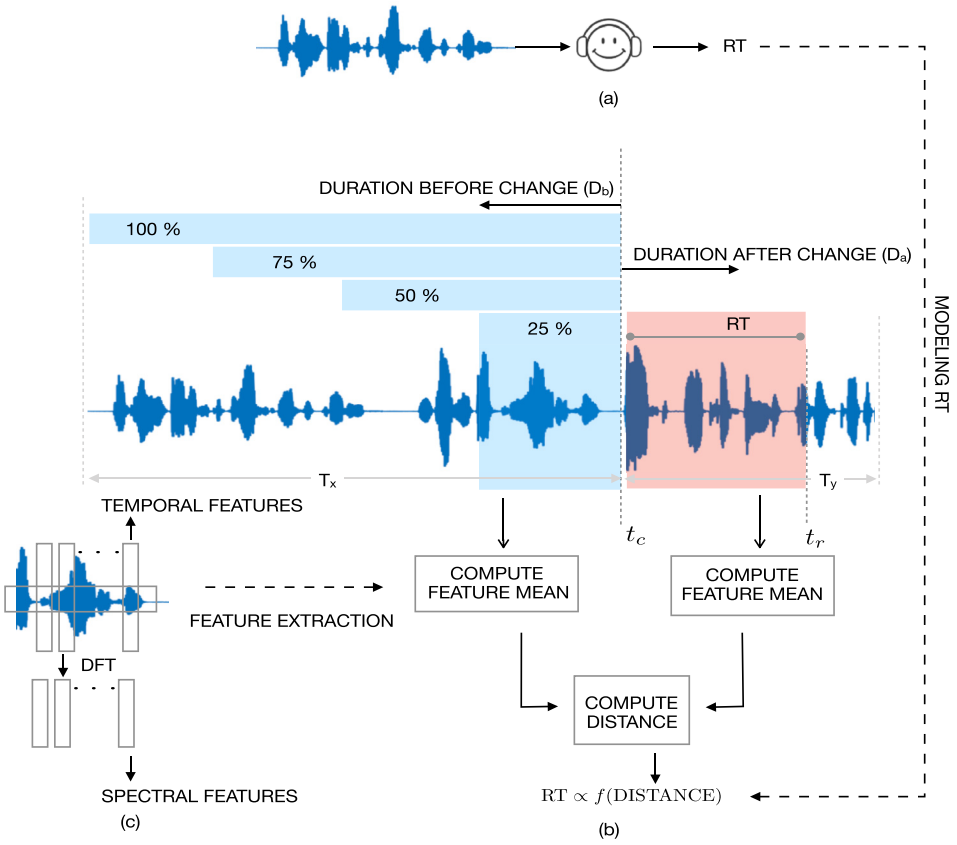

FIG. 7. (Color online) Proposed approach to model RT using acoustic features before and after change instant.

$$\underbrace{\begin{bmatrix} \log \mathrm{RT}_1 \\ \log \mathrm{RT}_2 \\ \vdots \\ \log \mathrm{RT}_{N_p} \end{bmatrix}}_{\mathbf{r}} = \underbrace{\begin{bmatrix} 1 & d_{k,1} \\ 1 & d_{k,2} \\ \vdots & \\ 1 & d_{k,N_p} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} w_0 \\ w_k \end{bmatrix}}_{\mathbf{w}}, \tag{1}$$

where w_0 and w_k are the model parameters representing the mean RT and slope of the regression line, respectively, and RT_i and $d_{k,i}$ denote the RT and feature distance in the ith trial, respectively. We solve for the model in Eq. (1) using minimum mean square error (MMSE). That is,

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} \|\mathbf{r} - \mathbf{D}\mathbf{w}\|_2 = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{r}, \tag{2}$$

$$\hat{\mathbf{r}} = \mathbf{D}\hat{\mathbf{w}}. \tag{3}$$

The MMSE optimized values of w_k for all subjects and for the nine features sets are shown in Fig. 8(a). These are obtained for D_b set to 100% of t_c. For a majority of subjects, w_k is negative for all the feature sets. This signifies that, on average, RT decreases with increased feature distance. Some feature sets have a greater negative slope than others. For example, the slope is maximally negative for MFCC-D and MFCC-DD feature sets. To quantify the modeling performance, we used the r-square measure,[53] computed as

$$r\text{-square} = 1 - \frac{\|\mathbf{r} - \hat{\mathbf{r}}\|_2^2}{\|\mathbf{r} - \bar{\mathbf{r}}\|_2^2}, \tag{4}$$

where $\bar{\mathbf{r}}$ is the mean of elements in $\mathbf{r}$. The r-square is also referred to as the "explained variance" by the model; a value close to 100% indicates good modeling performance; that is, the proposed model is able to better explain the observed variance in the data.

Figure 8(b) shows the obtained r-square (shown in % as explained variance) for different feature sets. For clarity of

TABLE I. Acoustic features used in the regression model analysis.

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-D	First-order temporal derivative of MFCCs	Spectral	12×1	25 ms
MFCC-DD	Second-order temporal derivative of MFCCs	Spectral	12×1	25 ms
ENGY-D	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

presentation, the depicted percentage is the average percentage across all subjects. At the level of individual features, the MFCC-D outperforms all other feature sets. Moreover, a majority of feature sets fall below 10%, thereby failing to explain a significant portion of the variance in RT.

To examine combinations of features sets, we performed a multiple linear regression by combining the feature distances from all feature sets as follows:

$$\underbrace{\begin{bmatrix} \log \mathrm{RT}_1 \\ \log \mathrm{RT}_2 \\ \vdots \\ \log \mathrm{RT}_{N_p} \end{bmatrix}}_{\mathbf{r}} = \underbrace{\begin{bmatrix} 1 & d_{1,1} & \dots & d_{9,1} \\ 1 & d_{1,2} & \dots & d_{9,2} \\ \vdots & & & \\ 1 & d_{1,N_p} & \dots & d_{9,N_p} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_9 \end{bmatrix}}_{\mathbf{w}}. \tag{5}$$

This gave the best r-square [see Fig. 8(b)]. A p-value (from hypothesis testing) illustration depicting the relevance of each feature set in the multiple linear regression is shown in Fig. 8(c). The MFCC-D and MFCC-DD are most relevant, $p < 0.05$, across all subjects. A scatter plot of RT versus estimated RT, pooling all subjects' trials, is shown in Fig. 8(d). The Pearson correlation coefficient amounted to 0.8, indicating a good overall estimation accuracy. Also, it can be seen that a major portion of true RT fall in the range 400–1000 ms, with few trials with RT > 1000 ms. In this range, the estimation is also concentrated along the $y = x$ line.

We also tested the modeling performance with decreasing duration of segment D_b, that is, duration set to 75%, 50%, or 25% of t_c duration before the change instant [as shown in Fig. 7(b)]. The result shown in Fig. 8(b) depicts that using 100% of the segment duration better models human performance, with systematic decreases as the sample of the first talker's speech decreases in duration.

D. Talker pair-wise analysis

To analyze the variability in TCD performance across the talker pairs, we examined talker-wise performance in TCD RT (see Fig. 9). Most of the talker pairs have the average RT (computed across subjects) in the same range, except $T_1 - T_5$ and $T_5 - T_1$. Also, the miss rate was found to be higher for these pairs of talkers. This suggests that these pairs may be overlapping a lot in the perceived talker space. Comparing the FA rate, averaged across subjects, talker T_3 had the highest FA rate.

IV. MACHINE SYSTEM FOR TCD

We evaluated the performance of three machine TCD systems on the same stimulus materials used in the human TCD experiment. The first system was an adaptation of a state-of-the-art diarization system, designed to segment audio into distinct talker segments based on i-vector and probabilistic linear discriminant analysis (PLDA).[54] Subsequently, the talker change instants can be obtained as segment boundaries. Traditionally, these systems use the whole audio file. We refer to this mode of operation as offline TCD.

In contrast to offline machine systems, listeners in the human TCD task did not listen to the whole audio file.

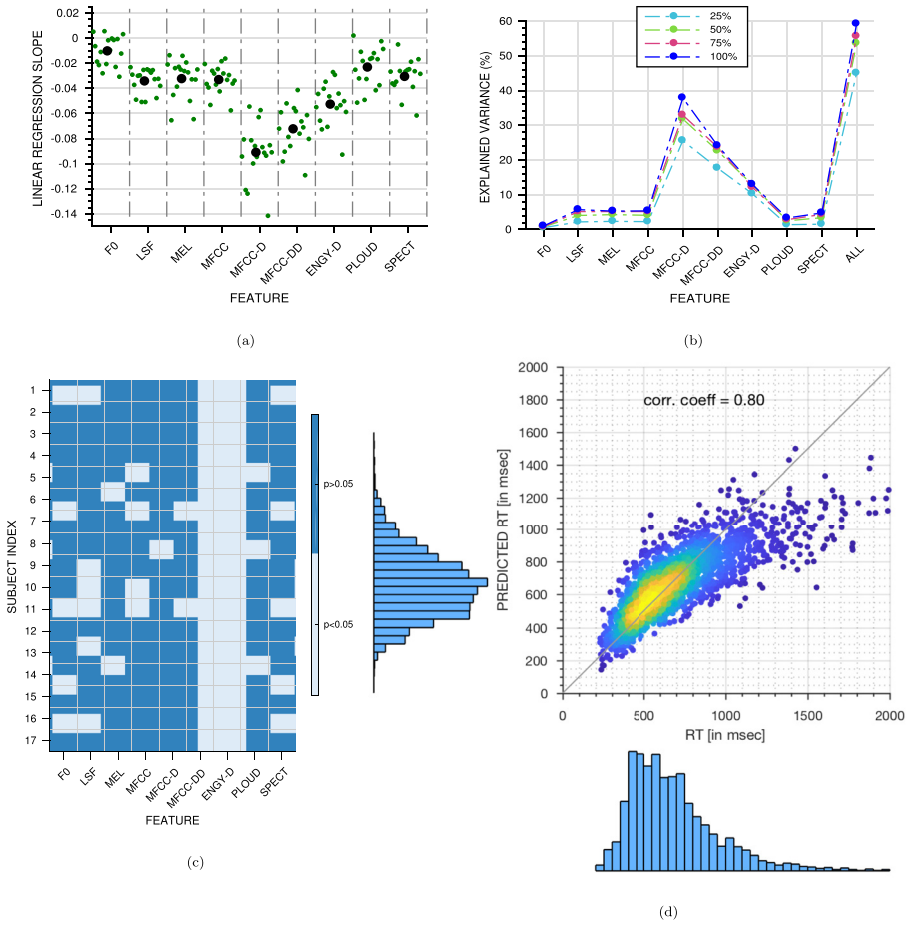

FIG. 8. (Color online) (a) Linear regression slope indicating the performance of individual feature sets in accounting for the variance in human TCD RTs. (b) The percentage of explained variance (r-square) for the linear regression model, averaged across subjects. (c) Illustration of significance (p-value) of different feature sets in multiple linear regression. (d) Scatter plot (polling all subjects' trials) of true RT versus estimated RT obtained from a multiple linear regression model computed across all feature sets.

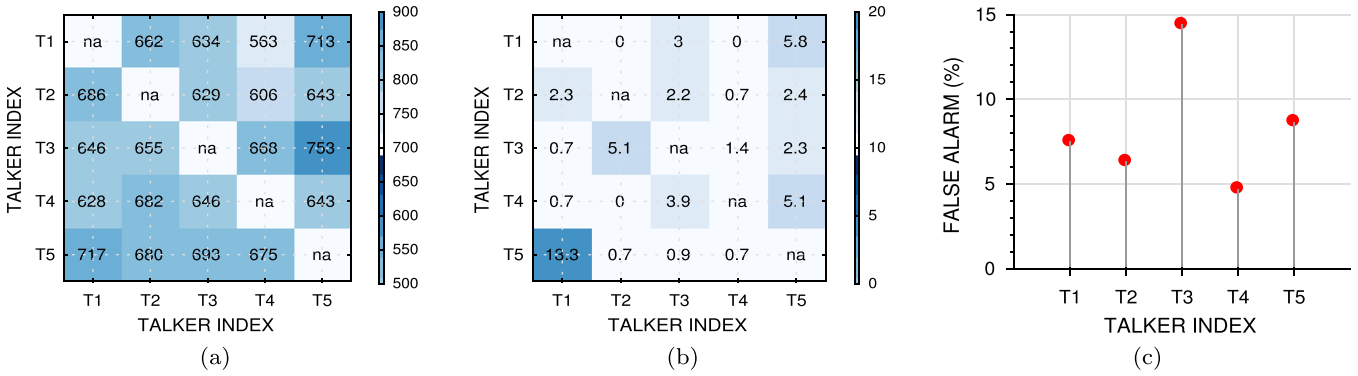

FIG. 9. (Color online) Dependence of (a) average RT on talker pairs (T_x-T_y), (b) average miss rate on talker pairs (T_x-T_y), and (c) average FA rate across talkers.

Instead, they performed an online change detection, pressing the button as soon as a talker change was perceived. To better model this aspect of human listening, we implemented an online variant of the offline machine system. Here, the system sequentially operated on segments of increasing duration starting with an initial 1 s segment for a given audio file. The sequential operation was stopped as soon as the second talker was detected, with this instant corresponding to the response instant for a talker change in machine recognition. An illustration of the approach is shown in Fig. 10. As we hypothesized for human listeners, the online system uses the acoustic features extracted from onset until the current segment duration in making a decision for talker segmentation.

The second machine system is the commercially available state-of-the-art IBM Watson Speech-to-Text (STT; https://console.bluemix.net/docs/services/speech-to-text/getting-started.html#gettingStarted) system that incorporates a talker change detector module.[55] The third system is purely based on textual features with no access to the acoustic detail of the speech. This text-based TCD system takes input from the transcript of the audio file and analyzes the semantic similarity between contiguous words for TCD. It thereby provides a control system with which to evaluate whether our approach to controlling semantic similarity in the novel TCD task (introducing a change in sentence context on both trials with and without a talker change) was successful.

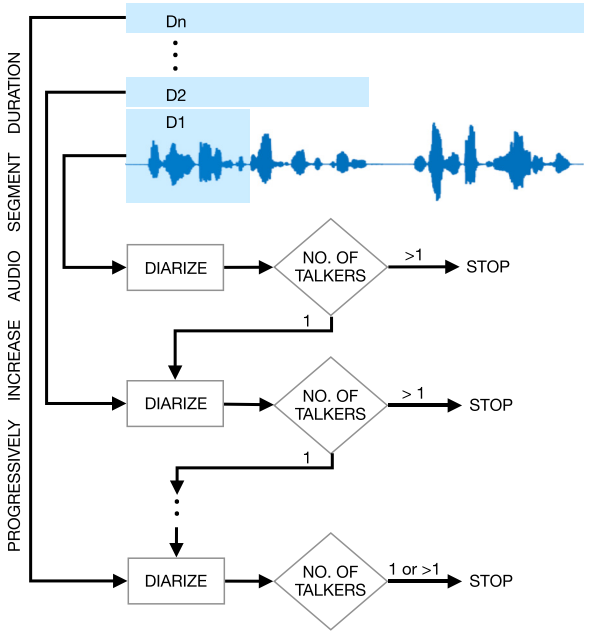

FIG. 10. (Color online) Proposed sequential diarization system as an online system for TCD.

For comparison of machine performance with human TCD behavior, the hit rate, miss rate, and the FA rate were computed using the same definitions as those for human TCD experiments. Sections IV A–IV C describe the systems and results.

A. Online and offline diarization-based TCD systems

We first segmented an audio file into small (750 ms) and temporally overlapping segments (temporal shift of 500 ms). Each short segment was transformed into an i-vector representation.[56] All pair-wise distances between the short-segment i-vectors were computed using a PLDA-based scoring approach. In the literature, PLDA-based scoring has been shown to impart robustness against small duration recordings (5–10 s) and variations in talking style,[57] and this suits the present stimulus set. Using the PLDA scores, agglomerative hierarchical clustering was performed to identify short-segments belonging to the same talkers and subsequently to merge the short-segments of the same talker. The obtained output was a segmentation of the input audio file into distinct single talker segments.

We developed the diarization system in the following two modes:

Offline diarization: Here, the diarization was performed on the complete audio file in one pass.
Online diarization: Here, instead of doing diarization across the complete audio file in one pass, we began with an input of 1 s and then sequentially increased it by 1 s, until two talker segments were detected or the end of file was reached (illustrated in Fig. 10).

The complete system setup was developed using the Kaldi toolkit.[58] This involved training the i-vector extractor 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.[43] 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 audio files corresponding to talkers used in the listening test were removed from the training dataset.

Since the extraction of i-vectors involves an overlapping shift, the system can flag a change instant before the actual ground-truth change instant. In order to account for this look-ahead, for these systems a tolerance window $\delta = 500$ ms was given and the RT corresponded to a hit if $t_c - \delta < \mathrm{RT} < t_c + 2000$ ms. The operating point for the system can be tuned by varying the segment clustering threshold. Thus, each operating point had its own miss- and false alarm rates. Hence, we generated a detection error trade-off (DET) curve for each system. These are shown in Fig. 12.

B. IBM Watson STT system

Along with an automatic speech recognizer (ASR), the commercially available IBM Watson STT system also includes a follow-up TCD system. Built as a one-of-a-kind approach,[59] the TCD system makes use of word boundaries from the ASR output in addition to acoustic feature information. For each acoustic segment corresponding to a pair of contiguous words, two separate Gaussian models are fit using MFCC features from left and right of the detected word boundary, respectively. A talker change is flagged based on the BIC algorithm or T^2 criterion. Use of word annotations from ASR reduces FAs in change detection within a word and in non-speech regions. We used the off-the-shelf implementation of this system available as a web application program interface. This system is pre-tuned and provides only one operating point. Like the diarization system, this system also was found to often give a change instant before the ground-truth change instant. This may be due to the larger temporal context used in the ASR module. On post analysis for hit rate computation, the detected change instant was found to lie in $t_c - \delta < \mathrm{RT} < t_c + 2000$ ms, with $\delta = 200$ ms. We considered all these responses as correct detections in hit rate computation. The resulting miss and FA rates are shown in Fig. 12.

C. TCD based on textual features

Although we hypothesize that acoustic feature distributions are likely to play a major role in the listeners' model of talker identity, it is also likely that the listeners attended to the semantics of the words making the stimuli.

Designing an experimental paradigm that mitigates semantic contributions in order to evaluate the acoustic features that contribute to TCD is challenging. The present paradigm takes the approach of introducing a *semantic* change on every trial. Hence, listeners cannot rely on a semantic change as a reliable cue for TCD as some trials feature no talker change. The challenge to disentangle the contribution of text/content versus acoustic features led us to examine a machine equivalent of TCD reliant upon only the information conveyed by the text of our stimulus material.[60,61] The proposed text-based TCD system is shown in Fig. 11. Using the transcripts from the LibriSpeech corpus, we trained a Word2Vec model with the Gensim Python package.[62] The model represents every word with 128-dimensional vector. This allows representing a sentence as a sequence of vectors obtained from the constituent words. We built a talker change detector by analyzing the semantic similarity among sets of words via analysis of the vector representations.

Specifically, the system is a classifier that takes N consecutive words as input and outputs a class label C_0 if all words are by a single talker and a class label C_1 otherwise. The value of N in this experiment was chosen after analyzing the average number of words in a sentence spoken by the first talker, as well as the average number of words spoken by the second talker within 2 s. Using features from 15 consecutive words from the corpus of talkers reading audio books, we generated training examples for the 2 classes. The training set corresponding to two different talkers (label C_1) was created by taking the last ten words of a sentence from the transcript of an audio book read by one talker and the first five words from the transcript of an audio book read by another talker. For the set of examples with no talker change (class C_0), the dataset was created by taking ten words from the end of one sentence and the first five words from the beginning of the following sentence, both drawn from transcripts of the audio book read by one talker. This dataset was created excluding the transcripts of audiobooks corresponding to the listening set. Using the training examples, we train a long short-term memory network (LSTM) to predict talker change based on context. The model was designed using the Keras toolkit[63] comprised of 1 LSTM layer with 512 cells followed by 3 dense layers with 1024, 512, and 256 neurons. The test corpus was made from the transcripts of the sentences used in the listening test. Each sentence was input as a series of $N = 15$ words to reflect the training set input, shifting a word by one to capture all words. A stimulus is marked as a change when it consists of at least one word from C_1. For computing DET curves, a threshold was used on the posteriors from the LSTM model. The resulting performance is shown in Fig. 12.

V. DISCUSSION

Summarizing the findings, the results from the experiments make three primary contributions toward understanding human and machine TCD performance.

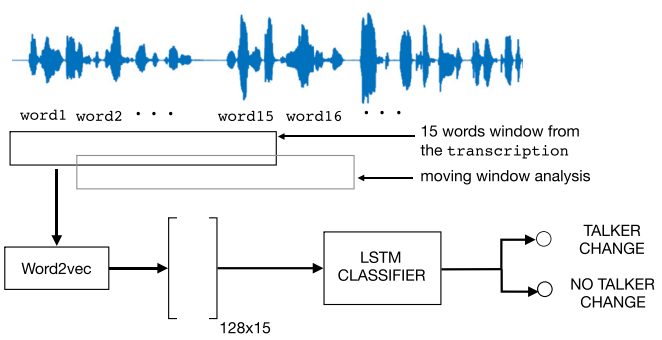

FIG. 11. (Color online) Illustration of text-based context change classification.

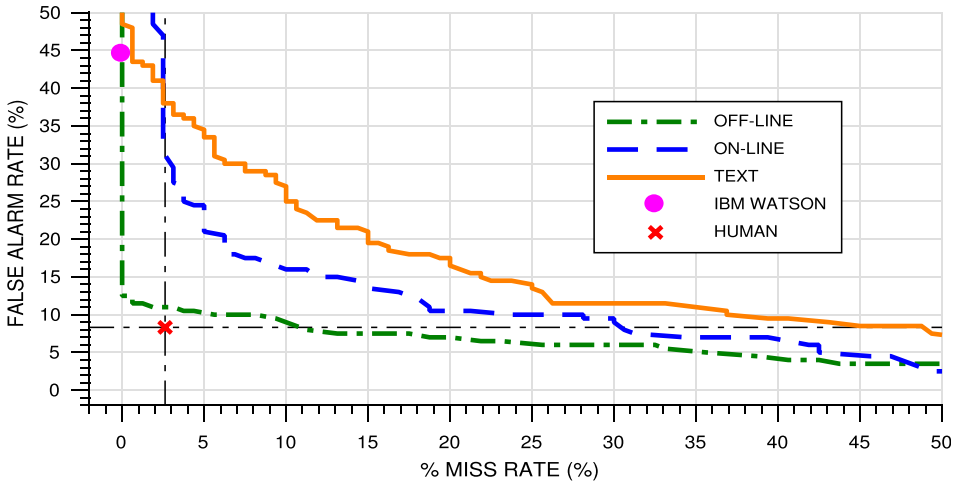

FIG. 12. (Color online) Detection error trade-off (DET) curve for all the TCD systems evaluated on the stimuli set used in the listening experiments. The IBM Watson system is a single operating point. The human score is also marked in this figure for reference.

A. Human performance for TCD

Human listeners performed the TCD task with very high accuracy, averaging only a 2.62% miss rate and 8.32% FA rate. The presence of FAs signifies that human listeners sometimes reported a talker change when there was none. This can be partly attributed to varying stress, intonation, and speaking rate exhibited by each talker during the course of reading the story book. At the same time, the high accuracy contrasts with some prior studies reporting high rates of "change deafness" to a change in talker.[11] However, it is important to note that the present task required listeners to direct the attention to detect a change in talker rather than to the comprehension alone. As suggested in prior research,[11] conversation expectations and task are likely to influence human listeners' direction of attention to fine-grain details of voice. At least in the context of an overt task requiring TCD, the present results demonstrate human listeners' ability to track acoustic features across talkers and differentiate the dimensions most relevant to detecting a change among male voices. Here, in the interest of investigating the issue in a controlled task, we examined read speech from audio books. The same approach might be applied to other stimulus sets capturing even more natural conversational speech, although in these cases covariation with context, semantic continuity, and other factors would be likely to complicate attempts to understand listeners' ability to track distributional information across acoustic dimensions that are related to talker identities.

Speculating from average RT to detect a change, the perceptual load for TCD appears to be greater than for simpler acoustic change detection scenarios such as noise-to-tone change or tone frequency change in the same listeners (experiments are reported in the supplementary material[45]). Pooling trials across listeners, the average RT for change detection was 680 ms (standard deviation = 274 ms), which was close to twice the average RT for a noise-to-tone change detection, and falls within the average RT associated with a tone frequency change detection task (done with different magnitudes of change). These differences may be attributed to recruitment of different auditory sub-processes for change detection in the context of speech and non-speech stimuli, specifically the need to accumulate distributional acoustic information across the more complex, multi-dimensional acoustic features that convey talker identity.

B. Estimating RT using a simple regression model on feature distances

We used a linear model to relate human listeners' log RT to detect a talker change and acoustic feature distances across a set of acoustic features. A simple Euclidean distance measure between the mean of the acoustic feature measurement corresponding to speech segment before and after change instant was used. Interestingly, we found the Piéron's law,[49] stating decrease in RT with increase in "strength of evidence" (such as loudness or frequency difference for tone stimuli), to hold for distance computed in the feature space [negative slopes in Fig. 8(a)] for TCD as well. Quantifying the model performance in terms of the percent of explained variance, the best fit was obtained in the MFCC-D feature space, followed by MFCC-DD and ENGY-D features. The 12-dimensional MFCC representation derived from MEL representation is a spectrally smoothed representation of the short-time spectrum, preserving the spectral peaks corresponding to formants with minimal pitch information. The MFCC-D representation, derived from the temporal derivative of MFCC, captures the rate of variation in the MEL representation. The improved model performance with the MFCC-D feature suggests that listeners are likely using the spectro-temporal variability in formant frequency space while attending to detect talker changes. We found a poor model fit with fundamental frequency ($F0$), often implicated in talker differences. This may be because there was considerable overlap in the fundamental frequency spread across our all-male set of talkers (shown in Fig. 2). Likewise, it is notable that application of a multiple regression model using the nine feature sets as independent variables improved the model performance (r-square ≈ 0.6). A significant contribution in this model came from MFCC-D and MFCC-DD [$p < 0.05$; see Fig. 8(c)]. This suggests that TCD involved tracking acoustic information across a quite complicated feature space. Visualizing the quality of estimated and true RTs (pooling from all subjects) we found a correlation of 0.8 [Fig. 8(d)]. A majority of true RTs fall in the range 400–900 ms, and a significant portion of this was mapped to the same range by the proposed model. Focusing

on the duration of speech segment before change instant used in the model estimation showed a systematic improvement in performance with duration for all the features [Fig. 8(c)]. This suggests that TCD response is less predictable using local information around the change instant and likely the listener continuously builds a talker model while attending to speech.

C. Machine performance on the TCD task

The offline diarization system provided better performance than the online system machine TCD system. This can be attributed to improved clustering of the i-vectors into two clusters because of availability of more data after change instant compared to the online case. The text-based TCD system exhibited relatively poor performance, indicating that, for the stimulus set, the text features did not suffice for TCD. This provides assurance that our novel paradigm for assessing human TCD (in which a change in sentence context was introduced on each trial independent of a talker change) was sufficient to reduce any bias arising from the semantic content of the utterance before and after the change instant. The IBM Watson has a FA rate too high for the successful use of the system in natural speech conversations, and underscores the importance of modeling the acoustic distributional characteristics of talkers' voices in supporting successful diarization. In fact, documentation of the system does caution that the presence of speaking style variations in the utterances may lead to high FA rates. Comparing human and machine systems we found a considerable gap in performance, with humans significantly outperforming state-of-the-art machine systems (see Fig. 12). In all, the present results highlight the complexity of the acoustic feature space that listeners must navigate in detecting talker change and underscores that machine systems have yet to incorporate the optimal feature set to model human behavior.

VI. CONCLUSIONS

The key contributions from the paper can be summarized as follows:

(i) Developing a novel paradigm for probing the human TCD across short-duration natural speech utterances.
(ii) Characterizing human TCD performance using various parameters—RT, hit rate, miss rate, and FA rate.
(iii) Building a simple linear regression-based model that estimates the human RT in TCD using the distance between mean acoustic features from speech segments. This model revealed the significance of MFCC-D and MFCC-DD acoustic features in TCD.
(iv) Comparing and benchmarking the machine TCD system performance implemented using principles of speaker diarization and textual features with human TCD performance.

ACKNOWLEDGMENT

The authors would like to acknowledge the generous support of Kris Gopalakrishan, BrainHub, and Carnegie Mellon Neuroscience Institute that fostered the collaboration between the Indian Institute of Science and the Carnegie Mellon University to pursue this work, Prachi Singh for the help with implementation of the machine systems, Purvi Agrawal and Shreyas Ramoji for several discussions, and all the volunteers who participated in this study.

[1]S. D. Goldinger, "Echoes of echoes? An episodic theory of lexical access," Psychol. Rev. **105**(2), 251–279 (1998).
[2]J. D. M. Laver, "Voice quality and indexical information," Br. J. Disord. Commun. **3**(1), 43–54 (1968).
[3]S. C. Levinson, "Turn-taking in human communication—Origins and implications for language processing," Trends Cognit. Sci. **20**(1), 6–14 (2016).
[4]L. C. Nygaard and D. B. Pisoni, "Talker-specific learning in speech perception," Percept. Psychophys. **60**(3), 355–376 (1998).
[5]P. T. Kitterick, P. J. Bailey, and A. Q. Summerfield, "Benefits of knowing who, where, and when in multi-talker listening," J. Acoust. Soc. Am. **127**(4), 2498–2508 (2010).
[6]I. S. Johnsrude, A. Mackey, H. Hakyemez, E. Alexander, H. P. Trang, and R. P. Carlyon, "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**(10), 1995–2004 (2013).
[7]M. J. Sjerps, H. Mitterer, and J. M. McQueen, "Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics," Neuropsychologia **49**(14), 3831–3846 (2011).
[8]Y. Lavner, I. Gath, and J. Rosenhouse, "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," Speech Commun. **30**(1), 9–26 (2000).
[9]G. Sell, C. Suied, M. Elhilali, and S. Shamma, "Perceptual susceptibility to acoustic manipulations in speaker discrimination," J. Acoust. Soc. Am. **137**(2), 911–922 (2015).
[10]T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**(2), 887–906 (2005).
[11]K. M. Fenn, H. Shintel, A. S. Atkins, J. I. Skipper, V. C. Bond, and H. C. Nusbaum, "When less is heard than meets the ear: Change deafness in a telephone conversation," Quart. J. Exp. Psychol. **64**(7), 1442–1456 (2011).
[12]M. S. Vitevitch, "Change deafness: The inability to detect changes between two voices," J. Exp Psychol: Human Percept Perform **29**(2), 333–342 (2003).
[13]J. G. Neuhoff, S. A. Schott, A. J. Kropf, and E. M. Neuhoff, "Familiarity, expertise, and change detection: Change deafness is worse in your native language," Perception **43**(2–3), 219–222 (2014).
[14]D. A. Coker and J. Burgoon, "The nature of conversational involvement and nonverbal encoding patterns," Human Commun. Res. **13**(4), 463–494 (1987).
[15]J. Kreiman and D. Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley, New York, 2011).
[16]M. Latinus, P. McAleer, P. E. Bestelmeyer, and P. Belin, "Norm-based coding of voice identity in human auditory cortex," Curr. Biol. **23**(12), 1075–1080 (2013).
[17]L. E. Humes and J. B. Ahlstrom, "Relation between reaction time and loudness," J. Speech, Lang., Hear. Res. **27**(2), 306–310 (1984).
[18]J. Schlittenlacher, W. Ellermeier, and G. Avci, "Simple reaction time for broadband sounds compared to pure tones," Atten. Percept. Psychophys. **79**(2), 628–636 (2017).
[19]D. S. Emmerich, D. A. Fantini, and W. Ellermeier, "An investigation of the facilitation of simple auditory reaction time by predictable background stimuli," Percept. Psychophys. **45**(1), 66–70 (1989).
[20]C. Suied, P. Susini, and S. McAdams, "Evaluating warning sound urgency with reaction times," J. Exp. Psychol. Appl. **14**(3), 201 (2008).
[21]C. Suied, P. Susini, S. McAdams, and R. D. Patterson, "Why are natural sounds detected faster than pips?," J. Acoust. Soc. Am. **127**(3), EL105–EL110 (2010).
[22]Y. Boubenec, J. Lawlor, U. Górska, S. Shamma, and B. Englitz, "Detecting changes in dynamic and complex acoustic environments," ELife **6**, e24910 (2017).
[23]E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Ninth European Conference on Speech Communication and Technology* (2005).

[24]J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," arXiv:1803.10609 (2018).
[25]G. Sell and A. McCree, "Multi-speaker conversations, cross-talk, and diarization for speaker recognition," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* (2017), pp. 5425–5429.
[26]O. Novotný, P. Matějka, O. Plchot, O. Glembek, L. Burget, and J. Černocký, "Analysis of speaker recognition systems in realistic scenarios of the SITW 2016 Challenge," in *Proc. INTERSPEECH, ISCA* (2016), pp. 828–832.
[27]X. Huang and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," IEEE Trans. Speech Audio Process. **1**(2), 150–157 (1993).
[28]A. G. Adam, S. S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* (2002), Vol. 4, pp. 3908–3911.
[29]J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," IEEE Signal Process. Lett. **11**(8), 649–651 (2004).
[30]N. Dhananjaya and B. Yegnanarayana, "Speaker change detection in casual conversations using excitation source features," Speech Commun. **50**(2), 153–161 (2008).
[31]V. Gupta, "Speaker change point detection using deep neural nets," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* (2015), pp. 4420–4424.
[32]R. Wang, M. Gu, L. Li, M. Xu, and T. F. Zheng, "Speaker segmentation using deep speaker vectors for fast speaker change scenarios," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* (2017), pp. 5420–5424.
[33]A. Tritschler and R. A. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in *Sixth European Conference on Speech Communication and Technology* (1999).
[34]M. Sarma, S. N. Gadre, B. D. Sarma, and S. R. M. Prasanna, "Speaker change detection using excitation source and vocal tract system information," in *2015 Twenty First National Conference on Communications (NCC)*, IEEE (2015), pp. 1–6.
[35]M. Yang, Y. Yang, and Z. Wu, "A pitch-based rapid speech segmentation for speaker indexing," in *Seventh IEEE International Symposium on Multimedia (ISM'05)* (2005).
[36]B. Abdolali and H. Sameti, "A novel method for speech segmentation based on speakers' characteristics," arXiv:1205.1794 (2012).
[37]W. N. Chan, T. Lee, N. Zheng, and H. Ouyang, "Use of vocal source features in speaker segmentation," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* (2006).
[38]H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* (1991), Vol. 2, pp. 873–876.
[39]S. S. Cheng, H. M. Wang, and H. C. Fu, "BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization," IEEE Trans. Audio, Speech Lang. Process. **18**(1), 141–157 (2010).
[40]A. S. Malegaonkar, A. M. Ariyaeeinia, and P. Sivakumaran, "Efficient speaker change detection using adapted Gaussian mixture models," IEEE Trans. Audio, Speech Lang. Process. **15**(6), 1859–1869 (2007).
[41]V. Karthik, D. Satish, and C. Sekhar, "Speaker change detection using support vector machine," in *Proc. 3rd Int. Conf. Non-Linear Speech Process* (2005), pp. 19–22.
[42]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.* (2015), pp. 5206–5210.
[43]H. Kawahara, I. Masuda-Katsuse, and A. de Cheveignè, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based *F*0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**(3-4), 187–207 (1999).
[44]https://gorilla.sc (Last viewed 15 August 2018).
[45]See supplementary material at https://doi.org/10.1121/1.5084044 for supplementary experiments and results on change detection.
[46]Sennheiser HD 215 II closed over-ear back headphone with high passive noise attenuation (Hanover, Lower Saxony, Germany).
[47]A. Mirzaei, S.-M. Khaligh-Razavi, M. Ghodrati, S. Zabbah, and R. Ebrahimpour, "Predicting the human reaction time based on natural image statistics in a rapid categorization task," Vision Res. **81**, 36–44 (2013).
[48]R. T. Pramod and S. P. Arun, "Do computational models differ systematically from human object perception?," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
[49]D. Pins and C. Bonnet, "On the relation between stimulus intensity and processing time: Piéron's law and choice reaction time," Percept. Psychophys. **58**(3), 390–400 (1996).
[50]L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition* (PTR Prentice Hall, Englewood Cliffs, 1993), Vol. 14.
[51]G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Tech. Rep., IRCAM (2004).
[52]B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *ISMIR* (2010), pp. 441–446.
[53]A. C. Cameron and F. A. G. Windmeijer, "An *R*-squared measure of goodness of fit for some common nonlinear regression models," J. Econometrics **77**(2), 329–342 (1997).
[54]G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Spoken Language Technology Workshop (SLT)*, IEEE (2014), pp. 413–417.
[55]https://github.com/IBM-Bluemix-Docs/speech-to-text (Last viewed August 4, 2018).
[56]N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE/ACM Trans. Audio, Speech Lang. Process. **19**(4), 788–798 (2011).
[57]I. Salmun, I. Opher, and I. Lapidot, "On the use of plda i-vector scoring for clustering short segments," in *Proc. Odyssey* (2016).
[58]D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, EPFL-CONF-192584 (2011).
[59]D. Dimitriadis and P. Fousek, "Developing on-line speaker diarization system," in *INTERSPEECH* (2017).
[60]Z. Meng, L. Mou, and Z. Jin, "Hierarchical RNN with static sentence-level attention for text-based speaker change detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), pp. 2203–2206.
[61]I. V. Serban and J. Pineau, "Text-based speaker identification for multi-participant open-domain dialogue systems," in *NIPS Workshop on Machine Learning for Spoken Language Understanding*, Montreal, Quebec, Canada (2015).
[62]R. Řehùøek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta (2010), pp. 45–50.
[63]F. Chollet, "Keras," available at https://keras.io (Last viewed 15 August 2018).