TARGET AND NON-TARGET SPEAKER DISCRIMINATION BY HUMANS AND MACHINES

Soo Jin Park†, Amber Afshan†, Jody Kreiman⋆, Gary Yeung† and Abeer Alwan†

†Dept. of Electrical and Computer Engineering, University of California Los Angeles, USA
⋆ Depts. of Head and Neck Surgery and Linguistics, University of California Los Angeles, USA

ABSTRACT

The manner in which acoustic features contribute to perceiving speaker identity remains unclear. In an attempt to better understand speaker perception, we investigated human and machine speaker discrimination with utterances shorter than 2 seconds. Sixty-five listeners performed a same vs. different task. Machine performance was estimated with i-vector/PLDA-based automatic speaker verification systems, one using mel-frequency cepstral coefficients (MFCCs) and the other using voice quality features (VQual2) inspired by a psychoacoustic model of voice quality. Machine performance was measured in terms of the detection and log-likelihood-ratio cost functions. Humans showed higher confidence for correct target decisions compared to correct non-target decisions, suggesting that they rely on different features and/or decision making strategies when identifying a single speaker compared to when distinguishing between speakers. For non-target trials, responses were highly correlated between humans and the VQual2-based system, especially when speakers were perceptually marked. Fusing human responses with an MFCC-based system improved performance over human-only or MFCC-only results, while fusing with the VQual2-based system did not. The study is a step towards understanding human speaker discrimination strategies and suggests that automatic systems might be able to supplement human decisions especially when speakers are marked.

Index Terms— Speaker perception, automatic speaker verification, voice quality, speaker discrimination

1. INTRODUCTION

Humans have a notable ability to distinguish individuals by their voices. Because perception studies suggest that distinguishing unfamiliar voices involves acoustic feature comparisons [1, 2, 3], we employed an *unfamiliar speaker discrimination* task to identify such features. In this task, listeners compared two speech samples to determine if they came from one speaker or from two different speakers. For automatic systems, speaker discrimination can be thought of as a special case of *speaker verification* where speakers are enrolled with one utterance.

Performance comparison between humans and machines on text-independent speaker discrimination tasks show that machines outperform humans on long utterances in certain conditions (e.g. [4, 5]). On very short utterances, however, machines seemingly perform worse than humans. For example, a state-of-the-art automatic speaker verification (ASV) system had an equal error rate of 22.31% with 2-sec-long pairs [6], while human listeners showed 11.4% miss and 19.7% false alarm rates for single sentence ($\approx$ 2 sec) pairs [7]. If humans are more accurate at this task, understanding perception might provide insights to improve machine performance.

Studies also suggest that speaker perception involves comparing voice tokens to stored prototypes in addition to featural comparisons [7, 8]. In the *prototype* model, humans encode the identity of a voice in terms of its deviations from an internal representation of a prototype or average voice. This model predicts that voices are more accurately identified if they are similar to the prototype than if they are dissimilar. Behavioral studies have provided evidence for such a model [9, 10]. However, acoustic features that characterize prototypes or the way in which they contribute to the model is not yet clear.

Automatic speaker discrimination can be viewed as analogous to the prototype model in that standard ASV systems build a universal background model (UBM, [11]) to represent an average speaker model, and the identity of a speaker is represented as a deviation from the UBM. Unlike perception studies, researchers can design a feature set to better represent speaker identity. In this sense, a reverse engineering approach relating human and ASV decisions might help develop a quantitative model for speaker perception.

Motivated by studies emphasizing the importance of voice quality for speaker perception [12], a voice quality feature set based on a psychoacoustic model [13, 14] was applied to ASV in our previous studies [15, 16]. This improved text-independent ASV system performance on very short-utterances ($\approx$ 2 sec). We then analyzed how humans and machines performed on a text-independent, short-utterance speaker discrimination task [17]. In that study, multidi-

We thank the Johns Hopkins University Human Language Technology Center of Excellence for providing the ASV system and computational resources. This research was supported in part by NIH and NSF.

mensional scaling was used to infer speaker spaces from the human and machine responses. Here, we further analyze the results of that study, focusing on the relationship between human and machine responses. Neurological data showed that speaker recognition and discrimination are separate abilities [1]. Considering that, we assume that perceptual strategies differ between target (same speaker pairs) and non-target trials (different speaker pairs). In this context, we relate responses by humans and machines for target and non-target trials, separately.

2. DATABASES

2.1. UCLA Speaker Variability Database

The UCLA Speaker Variability Database [18, 19] includes speech samples from 103 female and 105 male speakers, reflecting ordinary variations in voice quality due to multiple recording sessions, phonetic content, speaking style, and affect conditions. The speakers were recorded on 3 different days in a sound-attenuated booth, with a sampling rate of 22 kHz and a fixed mouth-to-microphone distance.

This study chose read sentences among the speech tasks in the database to represent the most stable and least varying type of continuous speech. Fifty female self-reported native speakers of English were randomly selected from the database. *Post hoc* listening by two linguists indicated that utterances from 9 speakers were perceptually "marked" by a non-American dialect (5 speakers), overly-precise articulation, and/or unusual dysfluencies in reading (4 speakers). The remaining 41 speakers lacked such personal idiosyncrasies, and are referred to as "unmarked".

2.2. NIST Speaker Recognition Evaluation Database

While the UCLA Speaker Variability Database provided all the evaluation utterances for the present study, separate speech databases were used to train the ASV systems tested here. The speaker recognition evaluation (SRE) databases developed by the National Institute of Standards and Technology (NIST) are often used to train a UBM and speaker variability subspaces; we used the SRE04, 05, 06, and 08 databases for this purpose [20, 21, 22]. Since the evaluation utterances were all from female speakers, only the recordings from female speakers were used for training. In addition, evaluation recordings were downsampled to an 8 kHz sampling rate to match the bandwidth of the SRE databases.

3. METHOD

3.1. Perceptual Speaker Discrimination

For each speaker, three read sentences ($<$ 2 sec each) were selected from each of the 3 recording sessions. These stimuli were assembled into 50 pairs of speakers in which both speech samples came from the same speaker and 1,225 pairs where the two speakers were different, for a total of 1,275 pairs. Stimuli were always drawn from different recording sessions, and two different sentences were used. Thus, this task is always text- and recording session-mismatched.

To minimize listener fatigue, stimuli were divided at random into 13 subsets. Thirteen groups of 5 normal-hearing subjects listened to the pairs of stimuli at a comfortable listening level. Each pair could be played only once in each presentation order (AB/BA). The listeners were asked whether the two speech samples were produced by the same speaker or by two different speakers. They also reported their confidence in their response on a 1–5 scale (1 = positive, 5 = wild guess). They were not told how many speakers were represented in the trials. The experiment was self-paced, and listeners were encouraged to take breaks as needed. Total testing time was less than one hour. For more details about the perceptual and automatic speaker discrimination experiments, see [17].

3.2. Automatic Speaker Discrimination

An i-vector [23]/PLDA [24] based ASV system was used to assess machine performance. The i-vector dimension was 600 and it was reduced to 200 after PLDA. The UBM was modeled with 2,048 Gaussian mixtures. The same stimuli presented to the human listeners were given to the ASV system to ensure a fair comparison.

Two feature sets were used in the experiments. The first was composed of mel-frequency cepstral coefficients (MFCCs) of dimension 20, along with their first derivatives. Second derivatives were not used because they did not provide notable performance gains in our preliminary work. The second feature set was inspired by a psychoacoustic model of voice quality [13, 14]. In a previous study [15], we tested the effectiveness of this feature set, after which the set was modified to better represent speaker identity for ASV [16]. The modified feature set (denoted as VQual2), included F0, F1, F2, F3, harmonic amplitude differences H1-H2, H2-H4, H4-H2k, formant amplitudes A1, A2, A3, and cepstral peak prominence (CPP, [25]). Here, H1, H2, H4, and H2k indicate the amplitudes of first, second, and fourth harmonics, and the harmonic nearest to 2 kHz. All features were automatically extracted without manual refinements.

3.3. Evaluation Metric

For humans, the similarity between the stimuli in each pair was measured by unfolding the confidence ratings such that a value of 10 (positive that voices are the same) meant the voices were very similar, and a value of 1 (positive that voices are different) meant they were maximally dissimilar. These scores were averaged across listeners. For ASV systems, the PLDA score, which represents the ratio of the likelihood that the given pair of stimuli are from the same speaker to the likelihood that the pair is from two different speakers, was used.

After obtaining the similarity scores from humans and PLDA scores from each automatic system, the scores were calibrated using standard logistic regression [26]. The resulting calibrated log-likelihood-ratio (LLR; L) represents the scalar responses by humans and the automatic systems.

The detection cost function (C_{det}), commonly known as DCF, and the log-likelihood-ratio cost function (C_{llr}) were used for performance evaluation [27]. C_{det} is defined as the expected cost of detection errors. It is a measure of discrimination suitable for evaluating application-dependent performance. For our application, C_{det} was obtained with cost of misses set at 25 and cost of false alarms set at 1, as the ratio between target trials and non-target trials.

On the other hand, C_{llr} is defined as an integral over a spectrum of operating points of C_{det}. Thus, C_{llr} is an application-independent measure for evaluating soft decisions. It can be interpreted as a measure of loss of information, thus the lower the C_{llr}, the more the average information per trial (in bits) increases by applying the system. C_{llr} has an analytic solution as shown in [27]:

$$C_{\text{llr}}(L_t) = \frac{1}{2}\left(\sum_{t\in\text{tar}} \frac{\log_2(1+e^{-L_t})}{N_{\text{tar}}} + \sum_{t\in\text{non}} \frac{\log_2(1+e^{L_t})}{N_{\text{non}}}\right)$$

where L_t is the log-likelihood-ratio for trial t; and where 'tar' is a set of N_{tar} target trials and 'non' is a set of N_{non} non-target trials. The two normalized summation terms represent expectations of 'log costs' for target trials (first term) and for non-target trials (second term), respectively.

The Bosaris toolkit [28] was used to calibrate the raw scores and for calculating C_{det} and C_{llr}. As the data size analyzed was limited, and as the main purpose of the study was to analyze calibration-independent performance, the calibration was trained and used on the same dataset.

3.4. System Fusion

Systems were fused based on the logistic regression method [29] using the Bosaris toolkit [28]. The fusion trains combination weights to fuse multiple systems providing a calibrated set of log-likelihood ratios.

3.5. Speaker-Level Analysis

The L and C_{llr} values were analyzed on the speaker level. For each of the 50 speakers, the L_t values for the trials including that speaker were collected. Then, mean values of L_t for target and non-target trials were calculated separately, denoted as L^{tar} and L^{non}, respectively. If L^{tar} is large for a speaker, this indicates that the speaker has small within-speaker variability. Similarly, if L^{non} is large for a speaker, it indicates that the speaker has small between-speaker variability, and it is difficult for the system to distinguish her from others.

C_{llr} can be representative of the reliability of the L score. The lower the C_{llr}, the more reliable the system responses are for the speaker. $C_{\text{llr}}^{\text{tar}}$ and $C_{\text{llr}}^{\text{non}}$, at the speaker level, were calculated in a similar manner.

4. RESULTS AND DISCUSSION

4.1. Human and Machine Performance

Human and machine performances are summarized in Table 1. As expected, humans performed better than machines. For example, humans' C_{det} was as low as 0.273, while values for the MFCC-based system and VQual2-based system were 0.500 and 0.682, respectively. Humans performed even better than fusion of the two automatic systems, which had $C_{\text{det}} = 0.513$. In addition, fusing human responses with any automatic system improved performance, consistent with [30]. This trend was preserved with different false alarm costs, and for the C_{llr} values.

When fusion improved performance, it suggested complementarity among systems. When the MFCC and VQual2 were fused, the $C_{\text{llr}}^{\text{non}}$ decreased from 0.739 to 0.721, without changing $C_{\text{llr}}^{\text{tar}}$. On the other hand, when humans responses were fused with VQual2, the $C_{\text{llr}}^{\text{non}}$ was not affected while the $C_{\text{llr}}^{\text{tar}}$ slightly decreased from 0.417 to 0.405. MFCCs provided more complementary information to human responses than VQual2 features did; they reduced $C_{\text{llr}}^{\text{tar}}$ and $C_{\text{llr}}^{\text{non}}$ from 0.417 to 0.342 and from 0.434 to 0.368, respectively.

Note that the data set was not split into development and evaluation sets for fusion, which might have resulted in some overfitting. In the future, with more data, we will repeat these experiments to ensure that no overfitting occurs.

4.2. Log-Likelihood-Ratio Analysis

Speaker-level L^{tar} and L^{non} are shown in Fig. 1. For humans, the target trial distribution had a smaller variance compared to that of the ASV systems. Additionally, the L^{tar} and L^{non} distribution for humans were well-separated. This explains higher human accuracy compared to machines.

Table 1. ASV performance in terms of detection cost functions (C_{det}), log-likelihood-ratio cost (C_{llr}), log-likelihood-ratio cost for target trials ($C_{\text{llr}}^{\text{tar}}$), and log-likelihood-ratio cost for non-target trials ($C_{\text{llr}}^{\text{non}}$). The plus ('+') symbol indicates a fusion between the systems.

	C_{det}	C_{llr}	$C_{\text{llr}}^{\text{tar}}$	$C_{\text{llr}}^{\text{non}}$
MFCC (M)	0.500	0.737	0.736	0.739
VQual2 (V)	0.682	0.884	0.897	0.872
Human (H)	0.273	0.425	0.417	0.434
M+V	0.513	0.728	0.736	0.721
H+M	0.216	0.355	0.342	0.368
H+V	0.273	0.419	0.405	0.434
H+M+V	0.231	0.353	0.341	0.365

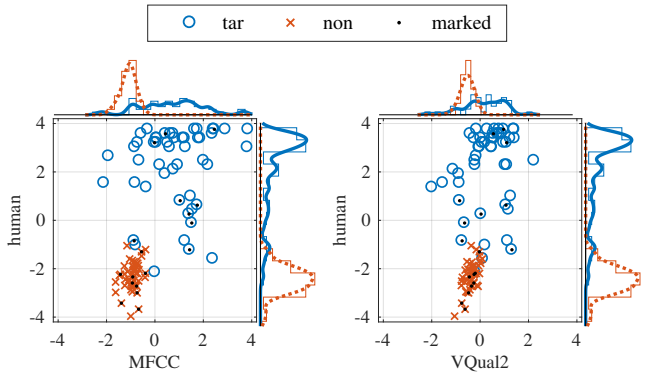

Fig. 1. Scatterplots of L^{tar} and L^{non} per speaker comparing MFCC vs humans (left) and VQual2 vs humans (right). L^{tar}s and L^{non}s are denoted with discs ('o') and crosses ('x'), respectively. Dots ('·') indicate perceptually-marked speakers.

Interestingly, the distributions of human responses were non-Gaussian and skewed towards correct responses. This tendency was more evident for the L^{tar} than L^{non}. That is, humans were more positive when they made "same speaker" responses than "different speaker" responses.

Next, the correlations between the L^{non} for humans and the two ASV systems were analyzed to understand which acoustic information was related to human responses (see Table 2). Compared to MFCCs, VQual2 had a high correlation with the human responses for L^{non} ($r = 0.610$). This suggests that human experts' decisions based on voice quality information could resolve false acceptances by the MFCC-based system [31]. Interestingly, for the 9 marked speakers, the correlation was even higher ($r = 0.912$). This might be related to findings that when linguistic cues are limited in the stimuli, human listeners assess speaker similarity of non-target pairs by relying on voice quality [32].

This tendency was not apparent for L^{tar}. Unfortunately, because only one target trial per speaker was made in this experiment, as opposed to 25 for the non-target trials, it is difficult to analyze what acoustic information was correlated with human responses for target trials.

4.3. Log-Likelihood-Cost Analysis

In our previous study, it was noted that human performance degraded when speakers were marked [17]. To analyze the relationship between speaker markedness and system reliability in terms of the information loss, C_{llr} values were analyzed.

For humans, the mean $C_{\text{llr}}^{\text{tar}}$ among the marked speakers was 0.784 compared to 0.417 for all speakers. For MFCCs, it was 0.574 among the marked speakers compared to 0.736 for all speakers. VQual2 showed no significant difference between the marked speakers and all speakers (0.909 and 0.897). That is, the MFCC system could take advantage of acoustic information for "same speaker" decisions from speaker markedness, while humans and VQual2-based systems could not. Moreover, when selecting the 4 monolingual English speakers among those marked speakers, human $C_{\text{llr}}^{\text{tar}}$ increased to 1.445. Those 4 speakers did not have non-American accents, but they had unusual dysfluencies in reading. The other 5 speakers' C_{llr} was much lower (0.256). This suggests that humans were not able to detect a consistent pattern in dysfluencies, but they could detect patterns for making "same speaker" decisions for the 5 speakers with non-American dialects. This hypothesis will be tested in future studies by including more target trials and marked speakers.

Table 2. Correlation coefficients of L^{tar} and L^{non} per speaker between each of the two ASV systems (MFCC and VQual2) and humans.

	MFCC	VQual2
L^{tar}	0.127	0.216
L^{non}	0.273	0.610

5. CONCLUSION

Speaker discrimination decisions by humans and machines on short-utterance, text-independent stimulus pairs were investigated. We focused on analyzing system responses and their reliability. System responses were measured in terms of the log-likelihood-ratio, and the reliability was calculated in terms of the log-likelihood-ratio cost function. Target and non-target trials were analyzed separately.

As expected, human listeners were considerably more accurate than machines. Higher confidence for correct target decisions compared to correct non-target decision was observed in the human response distribution. For non-target trials, system responses per speaker were highly correlated between humans and VQual2, especially when the speaker is perceptually marked. For target trials, humans response reliability decreased for marked speakers compared to when all the speakers were considered. However, MFCC response reliability was higher for marked speakers than all speakers. This suggests that MFCCs could extract information from speaker markedness for target trials, while VQual2 response reliability was not affected by speaker markedness. These results are consistent with the prototype model of speaker perception in that human decisions became less accurate with speaker markedness. Results additionally suggest that machines might be able to supplement human listeners in such conditions.

Future studies will include perception experiments with more target trials, as well as more marked speakers. In addition, male speakers will be studied for a gender-balanced analysis, and more detailed comparisons will be conducted with machines using other spectral and prosodic features.

6. REFERENCES

[1] D. Van Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[2] J. Kreiman and D. Sidtis, *Foundations of Voice Studies*, Wiley-Blackwell, 2011.

[3] Sarah V. Stevenage, "Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings," *Neuropsychologia*, 2018.

[4] V. Hautamäki, T. Kinnunen, M. Nosratighods, K. A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *Interspeech*, 2010, pp. 1473–1476.

[5] Juliette Kahn, Nicolas Audibert, Solange Rossato, and Jean Franois Bonastre, "Speaker verification by inexperienced and experienced listeners vs. speaker verification system," in *ICASSP*, 2011, pp. 5912–5915.

[6] R. K. Das, S. Jelil, and S. R. Mahadeva Prasanna, "Significance of constraining text in limited data text-independent speaker verification," in *SPCOM*, 2016, pp. 1–5.

[7] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Comm.*, vol. 10, no. 3, pp. 265–275, 1991.

[8] G. Papcun, J. Kreiman, and A. Davis, "Longterm memory for unfamiliar voices," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 913–925, 1989.

[9] S. R. Schweinberger, H. Kawahara, A. P. Simpson, V. G. Skuk, and R. Zäske, "Speaker perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.

[10] S. R. Mathias and K. von Kriegstein, "How do we recognise who is speaking?," *Frontiers in bioscience*, vol. 6, pp. 92–109, 2014.

[11] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech*, 1997, pp. 963–966.

[12] E. Gold and P. French, "An international investigation of forensic speaker comparison practices," in *ICPhS*, 2011, pp. 751–754.

[13] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, and Z. Zhang, "Toward a unified theory of voice production and perception," *Loquens*, vol. 1, no. 1, pp. 1–9, 2014.

[14] M. Garellek, R. Samlan, B. R. Gerratt, and J. Kreiman, "Modeling the voice source in terms of spectral slopes," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. 1404–1410, 2016.

[15] Soo Jin Park, Caroline Sigouin, Jody Kreiman, Patricia A. Keating, Jinxi Guo, Gary Yeung, Fang-Yu Kuo, and Abeer Alwan, "Speaker identity and voice quality: Modeling human responses and automatic speaker recognition," in *Interspeech*, 2016, pp. 1044–1048.

[16] Soo Jin Park, Gary Yeung, Jody Kreiman, Patricia A. Keating, and Abeer Alwan, "Using voice quality features to improve short-utterance, text-independent speaker verification systems," in *Interspeech*, 2017, pp. 1522–1526.

[17] Soo Jin Park, Gary Yeung, Neda Vesselinova, Jody Kreiman, Patricia A. Keating, and Abeer Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. 375–386, 2018.

[18] Jody Kreiman, Soo Jin Park, Patricia A. Keating, and Abeer Alwan, "The relationship between acoustic and perceived intraspeaker variability in voice quality," in *Interspeech*, 2015, pp. 2357–2360.

[19] Patricia A Keating, Jody Kreiman, and Abeer Alwan, "The UCLA Speaker Variability Database," 2018.

[20] A. F. Martin and C. S. Greenberg, "NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Interspeech*, 2009, pp. 2579–2582.

[21] M. Przybocki, A. Martin, and A. Le, "NIST speaker recognition evaluation chronicles - Part 2," in *Odyssey*, 2006, pp. 1–6.

[22] M. Przybocki and A. Martin, "NIST speaker recognition evaluation chronicles," in *Odyssey*, 2004, pp. 12–22.

[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *ICASSP*, vol. 19, no. 4, pp. 788–798, 2011.

[24] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *ICASSP*, 2013, pp. 7649–7653.

[25] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *JSLHR*, vol. 37, no. 4, pp. 769–778, 1994.

[26] P. McCullagh, "Generalized linear models," *European Journal of Operational Research*, vol. 16, no. 3, pp. 285–292, 1984.

[27] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I: Fundamentals, Features, and Methods*, pp. 330–353. Springer, 2007.

[28] Niko Brümmer and Edward De Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," 2011.

[29] Niko Brümmer, Luk Burget, Jan Honza Černocký, Ondej Glembek, Frantiek Grézl, Martin Karafiát, David A. Van Leeuwen, Pavel Matějka, Petr Schwarz, and Albert Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," in *ICASSP*, 2007.

[30] R. González Hautamäki, V. Hautamäki, P. Rajan, and T. Kinnunen, "Merging human and automatic system decisions to improve speaker recognition performance," in *Interspeech*, 2013, pp. 2519–2523.

[31] V. Hughes, P. Harrison, P. Foulkes, P. French, C. Kavanagh, and E. San Segundo, "Mapping across feature spaces in forensic voice comparison: The contribution of auditory-based voice quality to (semi-)automatic system testing," in *Interspeech*, 2017, pp. 3892–3896.

[32] E. San Segundo, P. Foulkes, and V. Hughes, "Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli," in *Australasian Conference on SST*, 2016, pp. 309–312.