A Comparison of Machine Learning Algorithms and Human Listeners in the Identification of Phonemic Contrasts

Paul Reid, Ksenia Gnevsheva, Hanna Suominen

Australian National University

reidp02@gmail.com, ksenia.gnevsheva@anu.edu.au, hanna.suominen@anu.edu.au

Abstract

To elucidate the processes by which automatic speech recognition (ASR) architectures reach transcription decisions, our study compared human and ASR responses to stimuli with manipulated cues for stop manner (burst, silence, and vocalic onset) and voicing (voice onset time, aspiration amplitude, and vocalic onset). Fourteen participants and two ASR systems completed a forced-response identification task. Results indicated that the cues were of perceptual significance for human participants, and though weighted differently, significant predictors of ASR output. This demonstrated that ASR systems may be relying on the same key acoustic information as do human listeners for phonemic classification.
Index Terms: applied linguistics, English, evaluation studies, natural language processing, speech perception, speech recognition software

1. Introduction

Although an increasing body of studies compare human speech perception and automatic speech recognition (ASR) as a performance evaluation metric in ASR [1], [2], [3], [4], acoustic cues remain potentially under-utilised as a tool to explore phonemic representations in ASR systems in comparison to humans. Human listeners use voice onset time (VOT), vowel first formant transitions, and aspiration amplitude as voicing cues for word-initial stops in English [5], [6], [7], [8], [9]. In stop manner perception, the burst of noise accompanying stop release, a rising first formant transition, and silence cue stop manner [10], [11], [12], [13]. Human listeners exhibit the phenomena of trading relationships [14], categorical perception [15], and cue hierarchies [9].

This study compared the significance of acoustic cues for stop consonant voicing and manner in human and ASR perception with the following research questions:

1) How are acoustic cues for stop consonant voicing and manner used by human listeners?

2) How are acoustic cues for stop consonant voicing and manner used by ASR systems?

3) What are the differences between ASR systems and human listeners in the use of these acoustic cues?

2. Method

2.1. Stimuli

2.1.1. Base stimuli

Base stimuli were created, which were then manipulated to alter the value of the acoustic cues under investigation using PRAAT [16]. The recording was conducted by a speaker of Australian English, using a Rode NT1A microphone, recorded at 44.1 kHz in .wav format on a computer. The carrier phrase 'I am going to say X now' was repeated 30 times for each of the three words 'pat', 'bat', and 'stay'. 25 representative tokens for each target were selected and then annotated and modified.

To create the 'pat'-'bat' base stimuli, the vocalic portion of the 25 representative 'bat' stimuli was spliced with a representative region of 80 ms of aspiration from one selected 'pat' stimulus. Each of the 25 base stimuli was annotated into two sections: aspiration (80 ms) and vowel onset (100 ms).

To generate the 'stay'-'say' base stimuli, (1) a representative burst was selected from one of the 'stay' recordings and was inserted into each of the other representative 'stay' recordings in place of their burst; (2) the same 200 ms of silence, extracted from the audio recording between stimuli, was inserted between the fricative and the burst for each stimulus. Each of the 25 'stay'-'say' base stimuli was annotated into three sections: silence (200 ms), burst (20 ms), and vowel onset (50 ms).

2.1.2. Human perception experiment stimuli

The final stimuli for the human perception experiment were generated in five groups from one selected 'pat'-'bat' base and in four groups from one selected 'stay'-'say' base.

The first three groups of 'pat'-'bat' stimuli had the vowel onset removed. The first group had the amplitude of the aspiration unchanged at 100%; the second and third groups had the amplitude of the aspiration increased to 200% and reduced to 50%, respectively. Within these three sets, the length of the VOT was then varied between 0 and 45 ms, in 3 ms increments, resulting in 48 stimuli (16 per group). The last two groups of 'pat'-'bat' stimuli had the aspiration amplitude left unmodified (i.e., 100%). The first of these groups had VOT reduced to 30 ms; the other had VOT reduced to 45 ms. For both these sets of stimuli, the vowel onset was removed in 10 ms increments from the left until all 100 ms was removed, which resulted in a total of 22 stimuli.

The first two groups of 'stay'-'say' stimuli had the entire vowel onset portion removed, and the second group also had the burst entirely removed. For both the with-burst and burst-less stimuli, the silence was varied between 0 and 200 ms in 10 ms increments resulting in 42 stimuli (21 per group). The final two groups of 'stay'-'say' stimuli first had the burst removed entirely, then the first of the two groups had silence set to 30 ms, the other had silence set to 100 ms. For both groups, the 50 ms of vowel onset was removed in 5 ms intervals from the left, resulting in 22 stimuli (11 per group).

2.1.3. ASR stimuli

Two sets of stimuli were processed by the ASR systems; the first allowed for a broad, multivariate analysis of the decision

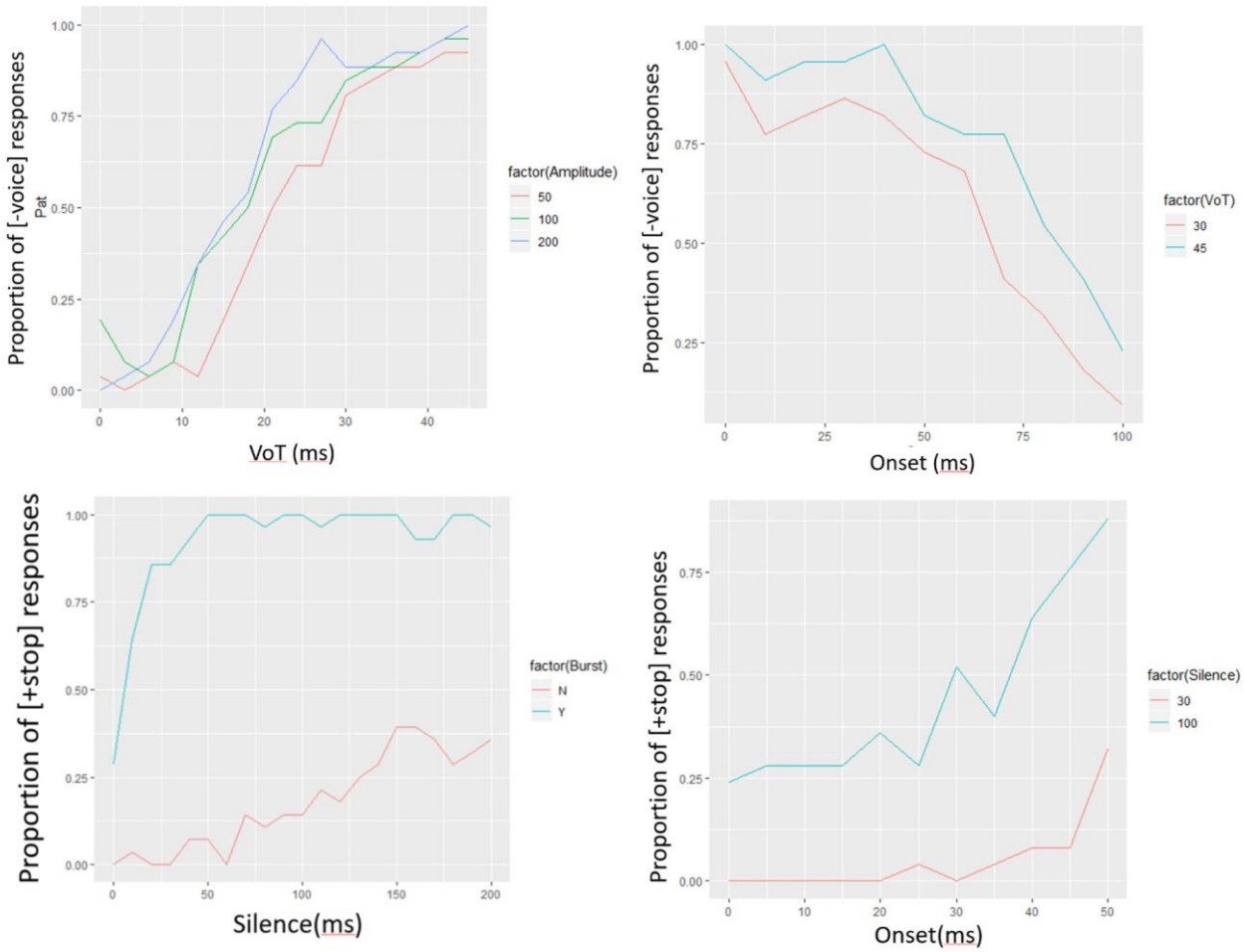

Figure 1: *Predictors of human participants' [-voice] and [+stop] responses.*

space of the ASR models (computer-only stimuli), whilst the second allowed a direct comparison between the ASR transcriptions and the responses in the human perception experiment (human comparison stimuli).

The computer-only 'pat'-'bat' stimuli were generated from each of the 25 base stimuli. The amplitude of the aspiration was varied between 50% and 200% of its original amplitude in 50% increments. For each of these sound files, the aspiration was cut in 10 ms increments to generate a new stimulus each time, starting from the right, until the aspiration was entirely removed. Finally, for each of those sound files, the vowel onset was cut in 10 ms increments from the left until the transition was entirely removed, which resulted in a total of 9900 stimuli.

The computer-only 'stay'-'say' stimuli were generated by modifying each of the 25 base stimuli. The vowel onset was cut in 5 ms increments, starting from the left, until the vowel onset section was entirely removed. For each of these, the burst was cut in 10 ms increments, starting from the right until the burst was entirely removed. Finally, the silence was reduced in 10 ms increments until entirely removed. This resulted in a total of 9075 individual 'stay'-'say' stimuli.

The acoustic manipulations for the human comparison stimuli were the same as the human perception experiment stimuli, carried out on all 25 base stimuli. This resulted in a total of 1,750 'pat'-'bat' stimuli and 1,600 'stay'-'say' stimuli.

2.2. Human perception experiment

The human perception experiment, which took approximately 30 minutes to complete, was conducted on a computer using the Qualtrics [17] interface. The participants were 14 monolingual speakers of Australian English between the ages of 20 and 40 (10 males and 4 females). The participants chose between 'pat' and 'bat' or 'stay' and 'say' after being presented with an audio stimulus of the form "I am going to say X now". The stimuli were placed into blocks by stimulus group, and the order of these blocks, the order of the individual stimuli within the blocks, and the order of the choices was randomized for each participant. This process was then repeated so that each block and stimulus was heard twice by each participant, which resulted in a total of 268 responses. Ethical approval (Protocol 2019/679) was obtained from the Human Research Ethics Committee of the Australian National University (ANU), and each study participant provided written informed consent.

2.3. ASR systems

The two ASR systems included in the experiments were Mozilla's DeepSpeech [18] and Google Speech-To-Text [19]. Mozilla's DeepSpeech is an open source, end-to-end ASR system based on the Recurrent Neural Network architecture. Version 0.6.0 pre-trained on American English was used for

ease of implementation. The default parameters were used when running the model, which was completed using a python script as per the documentation. Google's Speech-To-Text service is a cloud-based deep-learning based speech recognition system, which provides transcription services for 120 languages and dialects, including Australian English.

Non-target transcriptions represented no more than 3% of the total output for each set and were excluded from analysis. For each point in acoustic space represented by the values of the acoustic variables for the stimuli, the proportion of [-voice] and [+stop] output was calculated.

2.4. Analysis

Analysis of Variance (ANOVA) was conducted for both the human perception results and the output from the human comparison stimuli by the ASR systems. The independent variables tested were the acoustic variables modified for each group, as well as the interaction between them. Participant was used as an error term. Linear Discriminant Analysis (LDA) was used in the analysis of the computer-only output. An advisor from the ANU Statistical Consulting Unit was involved in this analysis design and its outcome reporting.

3. Results

3.1. Human listeners

Significant predictors of human listeners' ***voicing*** judgments in Group 1 stimuli were amplitude ($p < .001$) and VOT ($p < .001$): [-voice] responses increase with increase in VOT and amplitude (Figure 1, top left). Vowel onset ($p < .001$) and VOT ($p < .001$) were significant predictors of voicing in Group 2 stimuli: [-voice] responses decrease with longer onset and lower VOT (Figure 1, top right). Burst ($p < .001$), Silence ($p < .001$), and their interaction ($p < .05$) were significant predictors of ***manner of articulation*** judgments: in the with-burst condition [+stop] responses dominated, with a sharp drop at short silence; in the burst-less condition there was a slow, steady increase in the proportion of [+stop] responses (Figure 1, bottom left). Vowel Onset ($p < .001$), Silence ($p < .001$), and their interaction ($p < .001$) were significant predictors of [+stop] judgments: 100 ms silence was associated with a higher proportion of [+stop] responses, with a wide increase as vowel onset increased; [-stop] responses dominated at 30 ms silence with a slight increase at the longest onset values (Figure 1, bottom right).

3.2. ASR

The two ASR models demonstrated a distinct difference in ***voicing*** cues that were required to generate a [+stop] output, although this investigation was not the intent of the 'pat'-'bat' stimuli. DeepSpeech produced a very limited number of [+stop] transcriptions, instead outputting a fricative. Consequently, these results are excluded from analysis. For Google, as VOT increased, the likelihood of a [-voice] judgement also increased, except for VOT = 10 ms where [-voice] judgement dominated. As onset increased and amplitude decreased, so did the likelihood of a [+voice] judgement. LDA suggested that the three acoustic variables in combination were very powerful predictors of the model output. [+voice] and [-voice] were most widely separated by Vowel Onset and VOT, with the magnitude of the weighting for Amplitude (0.005) in Linear Discriminant 1 considerably lower than both Onset (0.02) and VOT (0.04).

The most obvious feature of the DeepSpeech [+stop] results related to the ***manner of articulation.*** Namely, the length of silence had a strong influence on the output, with an almost categorical [+stop] output for all values of Silence longer than 10 ms. As for Google, for Silence > 60 ms, [+stop] dominated regardless of the value of the other acoustic variables. The presence of the entire burst strongly cued [+stop]. At other Burst values, Onset remained a significant factor in determining the output until the limiting value of Silence was reached, which forced the [+stop] output.

LDA indicated that the three acoustic variables were strong predictors of the output for both models. For Google, LDA implied considerable separation for all three acoustic variables, whereas for DeepSpeech, Silence and Burst were well separated, whereas Onset much less so. For Google, LD1 weights of Silence (0.209), Onset (0.031), and Burst (0.017) were all relatively even while for DeepSpeech, Silence had the largest absolute LD1 weight (-0.053), followed by Burst (-0.01), and Onset (-0.003) was weighted very lightly.

3.3. Human vs ASR comparison

For Google, only the simple effect of VOT ($p < .001$) was shown to be significant for the human comparison stimuli for Group 1: low values of VOT are associated with predominantly [+voice] output (Figure 2, top left). VOT ($p < .001$) and Onset ($p < .001$), as well as their interaction, were significant for Group 2: [-voice] dominated at low values of Onset and decreased sharply at high values of Onset, with a higher value associated with a longer VOT (Figure 2, top right). The effect of Burst ($p < .001$) and the Silence:Burst interaction ($p < .001$) were significant for Group 3: in the with-burst condition, [+stop] was cued the majority of the time; in the burst-less condition, there was a steep rise from [-stop] to [+stop] output (Figure 2, bottom left). Silence ($p < .001$), Onset ($p < .001$), and the Silence:Onset interaction ($p < .001$) were significant for Group 4: longer silence resulted in categorical [+stop] responses while with shorter silence [+stop] increased as the vowel onset increased (Figure 2, bottom right).

For DeepSpeech, the simple effects of Burst ($p < .001$), Silence ($p < .001$) and the Burst:Silence interaction ($p < .001$) were significant for Group 3: [+stop] responses increased sharply at low silence duration, with a later rise in the burst-less condition. Onset ($p < .001$), Silence ($p < .001$) and the Onset:Silence interaction ($p < .001$) were significant for Group 4: [+stop] dominated for both Silence conditions, with a slight decrease in the long Silence condition at low Onset values.

4. Discussion

In human listeners, the significance of VOT [8], [9], vowel onset [5], [7], [8], and aspiration amplitude [9] in making stop voicing judgements in word initial position; and silence [10], [12], vowel onset [11], [13], and burst [13], [16] in making stop manner judgements in word medial position were confirmed. Our results supported claims that categorical effects are present in human perception of stop manner [20], [21] and stop voicing [22] contrasts. Previous research findings demonstrating a trading relationship between VOT and aspiration amplitude [9], and VOT and first formant [8] for stop voicing, and burst and silence [16] for stop manner were further supported. Finally, certain acoustic cues were only important within ambiguous values of more dominant cues, which supported the existence of a hierarchical relationship between the acoustic cues that were perceptually significant for a given phonemic contrast [9].

In ASR systems, for the stop manner contrast investigated, Burst, Onset, and Silence were all significant predictors for

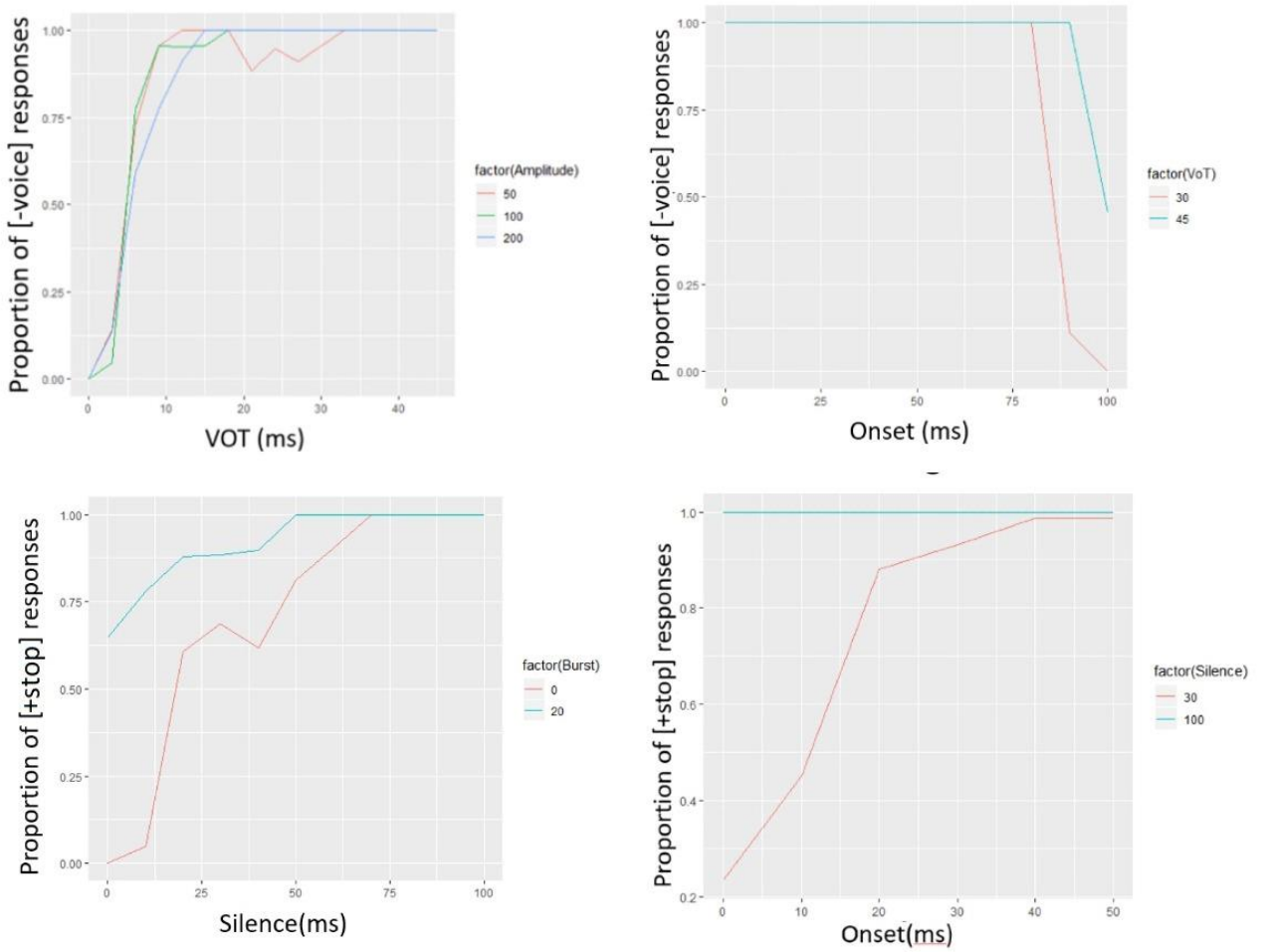

Figure 2: *Predictors of Google's ASR system responses.*

both systems; VOT, aspiration amplitude, and Onset were also significant predictors of stop voicing for Google. Further, a hierarchical relationship between the cues associated with a particular phonemic contrast and trading relationships between those acoustic cues were also present in ASR responses, in parallel with the phenomena observed in the human participants. The sharp gradient of the decision boundaries in response to changes in acoustic variables in stimuli also mirrored the categorical effect in humans.

The presence of these phenomena in both humans and ASR systems provided evidence to suggest that learned phonemic representations in the ASR systems may be based around the same kinds of acoustic information as humans use as their learning mechanisms. However, both experimented ASR models incompletely replicated the way in which humans make phonemic decisions, implying that the parallel between ASR and humans is not direct. In general, decision boundaries of the Google model appeared to more successfully replicate those from human perceptual results than DeepSpeech. Therefore, our results supported the claim that the sophistication of the phonemic representations of ASR systems is highly variable depending on the architecture used [23]. This finding highlighted the complexity of learning accomplished by ASR for making phonological judgements.

High-performance ASR systems, in particular neural network-based systems, underpinned by their methodological paradigm shift since 2010s explained by advancing deep learning, are capable of generating sophisticated representations of natural phonological classes [24], [25]. Since both ASR models, to varying degrees, used the same set of acoustic cues as humans to make phonemic decisions, this also implied that using these cues may be a natural consequence of optimising separation between categories based on the distribution of acoustic variables in that language. This was especially clear in the case of DeepSpeech, given its end-to-end architecture, since there was essentially no explicit acoustic engineering [26], and any phonetic cues learned to assist in the phonemic classification were based purely on distributional information within those phonemic categories.

If ASR systems develop comparable phonemic representations based on the same acoustic variables as humans, the theoretical ramifications could be significant. It may provide evidence for an auditory speech perception such as the Auditory Model [27], as it is unclear how an ASR system would represent phonemes in terms of articulatory gestures, as posited by Motor Theory [28]. It could also provide evidence for learning models based on the distributional properties of speech, such as statistical learning model [29]. It also presents a novel method for evaluating ASR performance by benchmarking against human perception; however, one must account for the spectral and temporal redundancies of acoustic cues [30] and covariance of acoustic cues in natural speech [31].

5. Acknowledgements

We thank the human participants for their time contribution. We acknowledge the support of the Australian Signals Directorate, the ASD-ANU Co-Lab, and Catherine Travis to the first author's thesis. We would also like to thank Associate Professor Alice Richardson from the ANU Statistical Consulting Unit for her statistical insights and expertise, which greatly assisted in this research.

6. References

[1] Deshmukh, N., Duncan, R. J., Ganapathiraju, A. and Picone, J. "Benchmarking human performance for continuous speech recognition", ICSLP'96 Fourth International Conference on Spoken Language Processing Proc., 1996.

[2] Goldwater, S., Jurafsky, D. and Manning, C. D. "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates", Speech Communication, 52(3):181-200, 2010.

[3] Kong, X., Choi, J.-Y. and Shattuck-Hufnagel, S. "Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures", IEE International Conference on Acoustics, Speech and Signal Processing, 2017.

[4] Richter, C., Feldman, N. H., Salgado, H. and Jansen, A. "Evaluating low-level speech features against human perceptual data", Transactions of the Association for Computational Linguistics, 5:425-440, 2017.

[5] Benkí, J. R. "Place of articulation and first formant transition pattern both affect perception of voicing in English", Journal of Phonetics, 29(1):1-22, 2001.

[6] Dmitrieva, O., Llanos, F., Shultz, A. A. and Francis, A. L. "Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English", Journal of Phonetics, 49:77-95, 2015.

[7] Liberman, A. M., Delattre, P. C. and Cooper, F. S. "Some cues for the distinction between voiced and voiceless stops in initial position", Language and Speech, 1(3):153-167, 1958.

[8] Lisker, L. "Is it VOT or a first-formant transition detector?", The Journal of the Acoustical Society of America, 57(6):1547-1551, 1975.

[9] Repp, B. H. "Relative Amplitude of Aspiration Noise as a Voicing Cue for Syllable-Initial Stop Consonants", Language and Speech, 22(2):173-189, 1979.

[10] Bastian, J., Delattre, P. and Liberman, A. M. "Silent interval as a cue for the distinction between stops and semivowels in medial position", The Journal of the Acoustical Society of America, 31(11):1568-1568, 1959.

[11] Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. and Gerstman, L. J. "Some experiments on the perception of synthetic speech sounds", Journal of the Acoustical Society of America, 24(6):597-606, 1952.

[12] Dorman, M. F., Raphael, L. J. and Liberman, A. M. "Some experiments on the sound of silence in phonetic perception", Journal of the Acoustical Society of America, 65(6):1518-1532, 1979.

[13] Liberman, A. M., Delattre, P. and Cooper, F. S. "The role of selected stimulus-variables in the perception of the unvoiced stop consonants", American Journal of Psychology, 497-516, 1952.

[14] Repp, B. H. "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception", Psychological Bulletin, 92(1):81, 1982.

[15] Kronrod, Y., Coppess, E. and Feldman, N. H. "A unified account of categorical effects in phonetic perception", Psychonomic Bulletin & Review, 23(6):1681-1712, 2016.

[16] Boersma, P. "Praat, a system for doing phonetics by computer", Glot International, 5:9/10: 341-345, 2001.

[17] Qualtrics. [Software]. http://www.qualtrics.com. 2017.

[18] Mozilla. Project DeepSpeech. Retrieved from https://github.com/mozilla/DeepSpeech 2020.

[19] Google. Speech-to-Text. Retrieved from https://cloud.google.com/speech-to-text/ 2020.

[20] Repp, B. H. "Closure Duration and Release Burst Amplitude Cues to Stop Consonant Manner and Place of Articulation", Language and Speech, 27(3):245-254, 1984.

[21] Bastian, J., Eimas, P. D. and Liberman, A. M. "Identification and discrimination of a phonemic contrast induced by silent interval", Journal of the Acoustical Society of America, 33(6):842-842, 1961.

[22] Harris, K. S., Bastian, J. and Liberman, A. M. "Mimicry and the perception of a phonemic contrast induced by silent interval: Electromyographic and acoustic measures", Journal of the Acoustical Society of America, 33(6):842-842, 1961

[23] Belinkov, Y. and Glass, J. "Analyzing hidden representations in end-to-end automatic speech recognition systems", Advances in Neural Information Processing Systems Conference, 2017.

[24] Nagamine, T., Seltzer, M. L. and Mesgarani, N. "Exploring how deep neural networks form phonemic categories", 16th Annual Conference of the International Speech Communication Association, 2015.

[25] Pellegrini, T. and Mouysset, S. "Inferring phonemic classes from CNN activation maps using clustering techniques", 17th Annual Conference of the International Speech Communication Association, 2016.

[26] Morais, R. "A Journey to <10% Word Error Rate". Retrieved from https://hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate/ 2017.

[27] Fant, G. Acoustic Theory of Speech Production, Mouton, 1960.

[28] Liberman, A. M. and Mattingly, I. G. "The motor theory of speech perception revised", Cognition, 21(1):1-36, 1985.

[29] Toscano, J. C. and McMurray, B. "Cue Integration with Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics", Cognitive Science, 34(3):434-464, 2010.

[30] Hermansky, H. "Coding and decoding of messages in human speech communication: Implications for machine recognition of speech", Speech Communication, 106:112-117, 2019.

[31] Whalen, D. H., Gick, B., Kumada, M. and Honda, K. "Cricothyroid activity in high and low vowels: exploring the automaticity of intrinsic F0", Journal of Phonetics, 27(2):125-142, 1999.