The Dawn of Psychoacoustic Reverse Correlation: A Data-Driven Methodology for Determining Fine Grained Perceptual Cues of Speech Clarity

Paige Tuttösí[1 2], H. Henny Yeung[3], Yue Wang[3], Jean-Julien Aucouturier[2], Angelica Lim[1]

Simon Fraser University, [1]School of Computing Science, [3]Department of Linguistics, Canada
[2]Université Marie et Louis Pasteur, SUPMICROTECH, CNRS, institut FEMTO-ST, France

ptuttosi@sfu.ca

Abstract

The production of clear speech has been extensively explored, and several contributing cues have been identified. However, synthesizing clear speech by mimicking these cues has shown poor results. We suggest that, rather than trying to replicate clear speech from produced human speech, we should instead use a data-driven approach to understand what cues are driving perception. In past work, we used psychoacoustic reverse correlation to show that vowel duration has a particularly important influence on the perception of English vowels among French adult learners of English. Here, we systematically controlled synthesized speech to identify duration patterns that bias a listener to a specific vowel. We find that increasing the duration of *tense* vowels improves clarity, but increasing the duration of *lax* vowels *reduces* the identification accuracy of those vowels. Moreover, we find that this mechanism is much stronger for those with reduced listening abilities, i.e., French learners of English. We hope that in the future a similar methodology can be used to explore these mechanisms for the hard of hearing.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Clear speech directed to those with comprehension challenges has been extensively studied from both a perception [1, 2, 3, 4, 2, 5, 6, 7], and production standpoint [8, 9, 10, 11]. Yet questions remain, specifically in terms of how to synthetically generate speech that is comprehensible for those with reduced listening abilities. When observing human-generated clear speech, one feature we see time and again is an increase in duration [8, 9, 2, 12], yet studies attempting to artificially incorporate these changes to mimic clear speech have not been particularly successful [4, 7, 5].

Past research had attempted to replicate human-made clear speech. While, in theory, we know this speech to be more intelligible, it is possible that the complex cues interacting in human speech are difficult to disentangle in efforts to replicate these changes in synthesized speech. As such, we propose that a bottom-up, data-driven approach focused purely on perceptive improvements is key to solving this problem.

In [13], reverse correlation allowed us to uncover psychoacoustic mechanisms that listeners use to differentiate sounds. Specifically, our previous results suggested a mechanism to improve comprehension when listeners struggle to use the primary vowel formant cues, which in our case was second language (L2) speakers [14], but this also applies to those with hearing loss [15]. However, we need to confirm that these mechanisms translate to macroscopic behavioural changes through validation experiments. We found a scissor-shaped effect where the duration within the target word should conform to the linguistic properties of the vowel, i.e., longer for tense and shorter for lax vowels, but a contrastive effect within the preceding context, i.e., faster preceding a tense vowel, and slower preceding a lax vowel. Still, we do not know the exact amount of duration change required to elicit the desired effects.

2. Experiment

2.1. Purpose

In this paper, to understand the effect of duration cues on clear speech, we aim to 1) establish perception thresholds by systematically manipulating the scissor manipulation's intensity, 2) additionally validate a new tense/lax vowel pair: "full" (/fʊl/) and "fool" (/ful/), and 3) test ecologically valid sentences, as our prior work on reverse correlation was conducted on ambiguous vowels (for methodological reasons, such as avoiding bias in a 1-interval task). Here, the vowels are no longer ambiguous.

2.2. Stimulus generation

We generated phrase stimuli in North American English (e.g. *"I heard them say fool"*) that incorporated the "scissor-shape" profile of speech-rate uncovered in [13]. To do this, we use Matcha-TTS [16], which has phoneme-level duration control. These modifications were made by applying an array of the same length as the phonemized phrase similar to the process in [17]. In this array, the phonemes up until the pause before the target word contained the context multiplier, and phonemes beginning at the space before the target word contained the word multiplier.

Within the word, we tested a duration change ranging from 0.5x speed to 2.0x speed at increments of 0.2, both increasing and decreasing from 1.0x speed, resulting in 11 different stretch manipulations. The corresponding context duration change ranged from 0.67x to 1.5x speed at increments of 0.1, both increasing and decreasing from 1.0x speed. Context and word duration were applied in opposite directions, e.g., when the word was 2.0x stretched, the context was compressed at 0.67x simultaneously. For each duration step, the phrases "I heard them say peel", "I heard them say pill", "I heard them say fool", and "I heard them say full" were generated, resulting in 44 different stimuli (4 phrases $\times$ 11 manipulations).

2.3. Experimental procedure

Participants were then presented successive trials consisting of a single stimulus, for which they were asked to choose which of two alternative words ("pill" or "peel", or "full" or "fool") they thought it included (1-interval, 2-alternative forced choice). Each of the 44 stimuli (4 phrases, 11 manipulations) were pre-

sented 5 times, in random order, resulting in 220 trials. The three conditions, word and context, word only, and context only, were varied between-participants. In each trial, participants could listen to the phrase once.

We had N = 75 French speakers to model L2 perception and N = 50, English L1 participants, with the latter group both in quiet and in noise to model a difficult listening environment for L1 speakers.

3. Results

3.1. Manipulation of both context and word

The results are presented as the percentage of correct identifications of the target words in the baseline, and the difference in percentage of correct identifications over the baseline.

French-L1. We found that a 1.6x stretch for tense vowels was required for a significant increase in transcription accuracy ($\geq 17\%$) over the baseline. Both tense and lax vowels required only a 1.2x compression to significantly decrease the accuracy over the baseline. Importantly, this shows that participants can be convinced to hear the tense vowel if the duration of a lax vowel becomes long (up to -37% for both pill and full). We observe that the hypothesized duration effects extend to our new vowel pair /ʊ/ and /u/. For "fool", specifically, the TTS struggled to synthesize clear formants and we see a strong improvement over the baseline (up to 22%) when extending the duration. As our sample of French speakers had high English proficiency, the baseline accuracy was relatively high now that the vowel was not ambiguous (80.8% for "pill", 76.8% for "peel", 83.2% for "full" and 32.0% for "fool"), yet were were still able to achieve the previously mentioned improvements.

English-L1. English L1 listeners were able to achieve near 100% baseline accuracy on all vowels, which differed from our prior work since vowels were no longer ambiguous and mostly insensitive to the manipulation. It is possible that because native speakers have more robust auditory representations of the spectral content of vowels in their native language, they are able to use the formants of the vowel to differentiate the words and are not easily swayed by changes in duration. The one exception was with "fool", where a 1.4x stretch resulted in a significant transcription improvement over the baseline ($\geq 20\%$). This is perhaps because "fool" was relatively poorly synthesized, suggesting that even English-L1 participants may start relying on duration cues in situations where timbral cues are not easily processed.

English-L1, with added noise. English-L1 results for "fool" stimuli, a word which, for technical reasons, was found to have poor synthesis results with our TTS system, suggest that, in more difficult or ambiguous listening conditions, even native speakers would default on secondary cues, similar to L2 speakers without tense/lax distinctions in their L1. To confirm this hypothesis and venture towards the possible application of our duration mechanism for L1 speakers with hearing loss, we explored whether it was possible to mask formant information with distortion and background noise, and this force L1 English speakers to behave like listeners with reduced comprehension for "peel" and "pill".

We then aimed to create a sound similar to loud-speakers in a metro station, i.e., a distorted and distanced sound, with a crowd in the background. To achieve this sound, we used Audacity[1]. First, a rectifier distortion was applied at 45%; then reverberation was added with a room size of 22%, a pre-delay of 10ms, a reverberance and dampening percentage of 50%, a tone low, tone high, and stereo width of 100%, and a wet and dry gain of -1bB. Finally, a background crowd noise was added so that all word duration in the phrase could be perceived, but it was difficult to recognize all of the speech clearly.

The L1 results in noise showed a similar effect as with L2 speakers. Duration manipulations affected word recognition performance both for "pill"(-22% with lengthening) and 'peel" (11% with lengthening, -52% with shortening). These results, therefore, suggest that the duration perception mechanism evidenced in reverse-correlation experiments is one used by both L1 and L2 speakers of English as a fall-back strategy when spectral information on the word itself is unreliable.

3.2. Manipulation of context-only and word-only

Finally, after confirming reverse-correlation results for simultaneous context and word manipulations of duration, we explored the contribution of manipulating only the context or only the word. We observed that, although reverse correlation kernels showed effects of duration both outside and inside the word, word identification performance was more strongly driven by word duration than context duration. The context manipulations resulted in only very slight improvements over the baseline, and modifying the word had as strong of an effect as modifying both the context and the word simultaneously.

3.3. Discussion

Clear speech literature often overlooks listener-specific differences, as well as the subtle differences in acoustics, often grouping them together [8]. The present work focuses on L2 perception and L1 perception in noise, however, we suggest that data-driven methodologies can also be applied to understand perception for specific listening groups, such as those with hearing loss. Moreover, although we focus on duration as in [13], future work will explore the other cues, through reverse correlation, for individuals with hearing loss.

4. Conclusions

Taken together, these results provide a promising strategy to improve the clarity of synthesized speech in difficult listening conditions, whether for a second language listener or a first language speaker in noise, by manipulating duration cues to enforce the correct perception of tense/lax alternatives. Specifically, because we did not see an improvement in either performance through the addition of duration changes in the context, we hypothesize that a word-only modification should be sufficient to improve comprehension in difficult listening conditions. Moreover, as the duration effects seen above were asymmetric, we make the hypothesis that lax vowels should remain at the base speaking rate, but words containing a tense vowel that can be easily confused with a lax vowel minimal pair should be lengthened relative to the rest of the phrase. This is contrary to previous work and studies of perception where all vowel sounds are lengthened to create clear speech [3, 12]. In follow up work currently in submission, we implement this strategy in a complete TTS system, using a parsing technique to automatically identify portions of the phrase that should benefit from durational changes and validate that this strategy improves speech comprehension compared to two other control strategies (slowing down the difficult word or slowing down the whole sentence). We propose our data-driven reverse correlation approach as a means to better understand clear speech.

[1] https://www.audacityteam.org/

5. References

[1] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1581–1592, 03 1994. [Online]. Available: https://doi.org/10.1121/1.408545

[2] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 259–271, 2002.

[3] W. V. Summers and M. R. Leek, "The role of spectral and temporal cues in vowel identification by listeners with impaired hearing," *Journal of speech and hearing research*, vol. 35, no. 5, pp. 1189–1199, 1992.

[4] Y. Nejime and B. C. J. Moore, "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 572–576, 01 1998. [Online]. Available: https://doi.org/10.1121/1.421123

[5] R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach, "Speaking clearly for the hard of hearing iv: Further studies of the role of speaking rate," *Journal of speech and hearing research*, vol. 39, no. 3, pp. 494–509, 1996.

[6] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3108–3118, 2004.

[7] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing. iii: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 32, no. 3, pp. 600–603, Sep. 1989.

[8] N. B. Aoki and G. Zellou, "Being clear about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for l1- and l2-english listeners," *Journal of Phonetics*, vol. 104, p. 101328, 2024.

[9] R. Scarborough and G. Zellou, "Clarity in communication: "clear" speech authenticity and lexical neighborhood density effects in speech production and perception," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3793–3807, 11 2013. [Online]. Available: https://doi.org/10.1121/1.4824120

[10] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[11] M. A. Knoll, M. Johnstone, and C. Blakely, "Can you hear me? acoustic modifications in speech directed to foreigners and hearing-impaired people," in *Interspeech 2015*, 2015, pp. 2987–2990.

[12] S. H. Ferguson and D. Kewley-Port, "Talker differences in clear and conversational speech: Acoustic characteristics of vowels," *Journal of speech, language, and hearing research*, vol. 50, no. 5, pp. 1241–1255, 2007.

[13] P. Tuttösí, H. H. Yeung, Y. Wang, F. Wang, G. Denis, J.-J. Aucouturier, and A. Lim, "Mmm whatcha say? uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation," in *Interspeech*, 2024, pp. 1010–1014.

[14] D. Kewley-Port, O.-S. Bohn, and K. Nishi, "The influence of different native language systems on vowel discrimination and identification," *The Journal of the Acoustical Society of America*, vol. 117, no. 4_Supplement, pp. 2399–2399, 2005.

[15] M. J. Hay-McCutcheon, N. R. Peterson, C. A. Rosado, and D. B. Pisoni, "Identification of acoustically similar and dissimilar vowels in profoundly deaf adults who use hearing aids and/or cochlear implants: Some preliminary findings," *American journal of audiology*, vol. 23, no. 1, pp. 57–70, 2014.

[16] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *ICASSP*, 2024.

[17] A. Joly, M. Nicolis, E. Peterova, A. Lombardi, A. Abbas, A. van Korlaar, A. Hussain, P. Sharma, A. Moinet, M. Lajszczak, P. Karanasou, A. Bonafonte, T. Drugman, and E. Sokolova, "Controllable emphasis with zero data for text-to-speech," in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 113–119.