Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes

Michelle Cohn[1] *and Georgia Zellou*[1]

[1]Phonetics Laboratory, Department of Linguistics, UC Davis, USA
{mdcohn, gzellou}@ucdavis.edu

Abstract

This study tests speech-in-noise perception and social ratings of speech produced by different text-to-speech (TTS) synthesis methods. We used identical speaker training datasets for a set of 4 voices (using AWS Polly TTS), generated using neural and concatenative TTS. In Experiment 1, listeners identified target words in semantically predictable and unpredictable sentences in concatenative and neural TTS at two noise levels (-3 dB, -6 dB SNR). Correct word identification was lower for neural TTS than for concatenative TTS, in the lower SNR, and for semantically unpredictable sentences. In Experiment 2, listeners rated the voices on 4 social attributes. Neural TTS was rated as more human-like, natural, likeable, and familiar than concatenative TTS. Furthermore, how natural listeners rated the neural TTS voice was positively related to their speech-in-noise accuracy. Together, these findings show that the TTS method influences both intelligibility and social judgments of speech — and that these patterns are linked. Overall, this work contributes to our understanding of the nexus of speech technology and human speech perception.

Index Terms: concatenative TTS, neural TTS, speech-in-noise perception, social ratings

1. Introduction

The recent pervasiveness of household voice-activated artificially intelligent (voice-AI) devices (e.g., Google Home, Amazon Echo) means that users are interacting with synthetic, text-to-speech (TTS) voices in their everyday lives. Yet, whether the speech generated by these modern systems is equally intelligible across different listening conditions (e.g., background talkers, music, fans, etc.) has not been thoroughly explored (cf., [1]). Further, there have been increasing efforts to make voice-AI speech as naturalistic as possible, resulting in more seamless, connected speech. For example, the application of long short-term (LSTM) neural networks in TTS (e.g., Wavenet [2]; for review see [3]) has resulted in more naturalistic connected speech that is rapidly being adopted industry-wide [4]. How differences in TTS methods impact speech intelligibility, however, is an open question.

In the present study, we consider how different TTS methods influence users' perception of synthesized speech. First, we test whether TTS generated via concatenative versus neural synthesis methods might result in differences in intelligibility during speech-in-noise perception. In concatenative TTS, individually recorded utterances are chunked into segments then re-combined via unit selection, which listeners perceive as having prosodic peculiarities (cf. [5]). Further, the concatenation process, particularly for pre-recorded real words, lacks the between-word coarticulation, or articulatory overlap, that is present in natural, connected speech [6]. Autoregressive neural TTS methods, on the other hand, generate words that are conditioned on all previous utterances, as well as on the immediately preceding segmental content (the local acoustic-phonetic context), resulting in significantly higher perceived 'naturalness' ratings by listeners [2]. This reported difference in perceived naturalness leads us to ask a second question: are there differences in how users rate social characteristics of the neural and concatenative TTS voices? Finally, we relate these two speech perception behaviors and test whether these social judgments related to intelligibility.

In the following sections, we provide a background on the speech-in-noise perception literature (§1.1), reviewing differences based on 'clear' and 'connected' speech (§1.1.1), the impact of semantic context (§1.1.2), and finally individual differences in listeners' ratings of social attributes and how they are related to speech intelligibility in the human-human literature (§1.1.3).

1.1. Speech-in-noise Perception

Difficulty in perceiving speech in the presence of noise is well attested in the literature [7]–[11], particularly for adults and children with hearing impairment [8]–[10]. Competing auditory signals (e.g., a lawn mower, other talkers, etc.) can interfere with a listener's ability to hear a speaker's intended message, whether they are a dinner companion in a noisy restaurant or if they are a voice-AI device in a noisy room. While work has shown that the type of masking noise has an effect on perception (e.g., multitalker babble versus white noise in [11]), the acoustic-phonetic properties of the speech signal are also a factor. For one, the type of voice matters: listeners show lower intelligibility for TTS voices, relative to naturally-produced human voices [12].

1.1.1. Casual vs. clear speech

Speech style also has an impact on perception: many studies have shown reduced intelligibility for more 'casual' and connected speech, relative to 'clear' speech [13]–[17]. Based on their synthesis differences, neural and concatenative TTS could serve as proxies for 'casual' and 'clear' speech, respectively. For one, neural TTS is more likely to contain phonetic reductions, typical of natural human speech; for example, [18] show that neural TTS automatically generated speech that included filled pauses (e.g., "um") after training on podcast episodes. Concatenative TTS, on the other hand, is more likely to result in relatively more hyper-articulated 'clear' speech; each segment is carefully selected and combined. Accordingly, we can set up several predictions for the present study. On the one hand, less effortful and casual speech results in shorter durations and less canonical segments than in speech produced in a clear manner ('clear speech' [19]). As a result, there is less robust content for listeners to glean from a noisy signal. Therefore, one prediction is that speech-in-noise perception for neural TTS utterances will be more difficult for listeners, relative to concatenative TTS, in line with prior work in the human-human literature [13]–[17].

On the other hand, greater coarticulatory overlap [6], or segmental connectedness, in casual speech might improve 'auditory streaming', allowing listeners to 'chunk' sets of sounds and disentangle them from background noise [20]. For example, [20] presented listeners with synthetic TTS varying in degree of coarticulation on vowel F2 (cueing neighboring /r/ or /z/); they found that speech-in-noise accuracy was higher when the TTS output included the consonant-vowel coarticulation, relative to when it did not. In the human-human literature, there is also some evidence that coarticulation can be helpful: [21] found that listeners displayed faster reaction times for words produced with greater coarticulation in a lexical decision experiment, than when the words were produced with less coarticulation. Accordingly, another prediction for the present study is that more connected TTS methods (i.e., neural TTS) might improve intelligibility across increasingly noisy contexts, with greater cues of coherence to extract a word from background noise.

1.1.2. Semantic predictability

In addition to acoustic-phonetic variations as a function of clarity (e.g., 'clear' versus 'casual' speech), listeners also use semantic context from an utterance to aid in word identification of speech in noise (e.g., [22]–[25]). How listeners integrate this context, however, may differ according to *how* their interlocutor is speaking: for example, [24] found worse keyword identification in semantically anomalous contexts, but less of a decline when the utterance was produced in 'clear' compared to 'casual' speech (produced by the same talker). In other words, the effect of semantic context on word intelligibility is mediated by the acoustic-phonetic properties of the utterance. Thus, we ask whether the effect of semantic predictability during speech-in-noise perception differs across concatenative and neural TTS. One prediction is that neural TTS, conditioned on the previous utterances (i.e., long-term) as well as the immediately preceding acoustic context (i.e., short-term), will improve intelligibility of the final target word since it provides more robust acoustic-phonetic cues in the signal (e.g., coarticulation) which listeners might be able to leverage when semantic context is not helpful. An alternative prediction is that neural TTS will result in even lower accuracy for low predictability sentences, if listeners are not able to disentangle the target utterance from the competing background noise.

1.1.3. Individual differences in speech-in-noise perception

There is also a great deal of variation among listeners in speech-in-noise tasks (cf. [26], [27]). For one, a listeners' familiarity with the speech variety has been shown to influence their speech-in-noise perception [22], [23], [28]. For example, [22] found that the intelligibility benefit of semantically predictable contexts is reduced when the speaker produces a dialect that the listener is unfamiliar with.

Others have shown differences in intelligibility and participants' 'likeability' of certain synthetic voices (but note that a direct relationship between TTS voice and these ratings was not observed) [29]. One possibility is that, in [29], there was a confound between the socio-indexical characteristics of the TTS voice itself and intelligibility. Here, we disentangle these factors by holding the 'speaker' constant (i.e., same set of AWS Polly voices), but manipulating the type of TTS method. Therefore, an additional consideration in the present study is whether individual language attitudes of the TTS voices may relate to their intelligibility under difficult listening conditions. In particular, the current study tests whether there are differences in how listeners perceive neural and concatenative TTS voices for four dimensions: how 1) machine-like / human-like, 2) unfamiliar / familiar, 3) eerie / natural, and 4) unlikeable / likeable the voice sounds. We predict that there will be a relationship between these ratings and intelligibility: in particular, that voices rated as more human-like, natural, and familiar will show intelligibility benefits, in line with the work on naturally produced voices.

1.2. Current Study

The present study consisted of two experiments. In Experiment 1, we test keyword identification accuracy of sentences presented in noise (comparing semantically predictable and unpredictable contexts) for speech generated from two different types of TTS methods: neural and concatenative TTS. Both TTS types were trained on 4 identical speaker datasets. Using TTS voices generated by distinct methods allows us to explicitly test predictions about the role of neural versus concatenative speech on intelligibility. It also provides a benefit for direct replication of this study in other labs, where idiosyncratic properties of recruited speakers may otherwise contribute to differences in their relative intelligibility. In Experiment 2, we collect each participant's ratings of four social attributes: human-likeness, familiarity, naturalness, and likeability of each voice. We first test whether there are systematic differences in these ratings by TTS Condition (neural vs. concatenative) and then relate patterns of variation directly to intelligibility ratings in Experiment 1.

2. Experiment 1: Intelligibility in Noise

2.1. Methods

Participants consisted of 28 native English speakers (24 female; mean age = 19.29 years, sd = 1.41 years) recruited through the UC Davis Psychology subjects pool. 26 participants reported that they had experience using at least one voice-AI system: 15 for Amazon Alexa, 8 for Google Assistant, and 11 for Apple's Siri.

We selected 192 sentences from the Speech Perception in Noise (SPIN) test [30], where monosyllabic target words occurred sentence-finally. Half of the sentences contained target words were semantically predictable based on context (e.g., "The boat sailed along the coast."), while the other half were semantically unpredictable (e.g., "Miss Brown might consider the coast."). Using AWS Polly, all 192 sentences were generated with both concatenative TTS and neural TTS for 4 adult female Amazon Polly voices (US-English): 'Salli', 'Kendra', 'Kimberly', and 'Joanna'. All sound files were resampled to the lower sampling rate of the two (neural TTS: 22,050 Hz) and amplitude-normalized (60 dB). The beginning and end of each sentence was padded with 800 ms of silence; as a result, when mixed with the speech-shaped noise, all target sentences were gated into noise. Next, we generated speech-shaped noise using the long-term average spectrum

(LTAS) for all sentences combined [31], [32]. Then, all sentences were combined with the speech-shaped noise at two signal-to-noise ratios: -3 dB and -6 dB SNR [33].

Participants completed the experiment in a sound-attenuated booth in the UC Davis Phonetics Lab. Participants were seated in the booth facing a computer monitor and keyboard and wearing over-ear headphones (Seinheiser Pro). On each trial, participants heard a sentence and were prompted to type the last word of the sentence using the keyboard. The 192 sentences were presented equally across the 8 voices (4 speakers x 2 TTS conditions) and 2 SNRs; which sentence was presented in which condition was fully randomized between subjects. Finally, participants completed a short hearing screen (250-8000 Hz [34]). Data for participants who did not pass the screening were excluded from the analysis.

2.2. Word Identification Analysis

Keyword accuracy on each trial was coded as binomial data (1 = correct word identification, 0 = incorrect) automatically using string matching. Trial accuracy (1 or 0) was modeled with a mixed effects logistic regression with the *lme4* R package [35]. Fixed effects included TTS Condition (2 levels: concatenative, neural), Signal-to-Noise Ratio (2 levels: -3 dB SNR, -6 dB SNR), Semantic Predictability (2 levels: low, high), and all possible interactions. Random effects included by-Listener random intercepts and by-Listener random slopes for TTS method, SNR, and Semantic Predictability conditions. Additionally, we included by-Talker random intercepts to account for variation in baseline intelligibility for each speaker dataset. Contrasts were sum coded.

2.3. Word Identification Results

Table 1 provides the summary statistics of the accuracy model. Figure 1 shows the mean accuracy across the conditions. First, there was a main effect of TTS Condition: listeners were less accurate at keyword identification for neural TTS than for concatenative TTS (see Figure 1). SNR level and Semantic Predictability were also significant main effects: listeners were less accurate at keyword identification for sentences presented at a lower SNR (-6 dB SNR), relative to a higher SNR; and listeners were less accurate at identifying words occurring in low semantic predictability sentences than high semantic predictability. There was also a significant interaction between SNR and Semantic Predictability: low predictability sentences at a low SNR (-6 dB SNR) had even lower keyword identification accuracy.

Table 1: *Model summary for word identification accuracy*

	Beta Coef	*Std Error*	*z*	*p*
(Intercept)	0.13	0.07	1.8	0.07
TTS(Neural)	-0.21	0.03	-6.2	<0.001***
SNR(-6)	-0.53	0.03	-16.5	<0.001***
Predict(Low)	-0.73	0.03	-23.7	<0.001***
TTS(Neural) x SNR(-6)	-0.04	0.03	-1.3	0.20
TTS(Neural) x Predict(Low)	-0.02	0.03	-0.5	0.59
SNR(-6) x Predict(Low)	0.01	0.03	4.3	<0.001***
TTS(Neural) x SNR(-6) x Predict(Low)	0.03	0.03	0.9	0.35

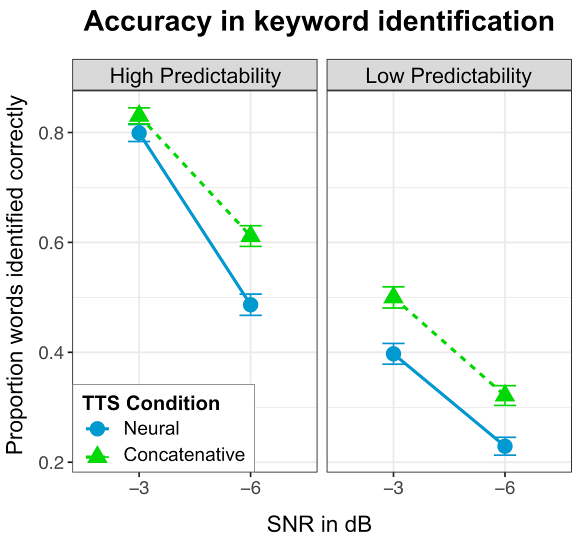

Figure 1: *(Experiment 1) Mean accuracy of keyword identification in high and low Semantic Predictability contexts, lower and higher SNR, across two TTS synthesis types. Error bars = standard error.*

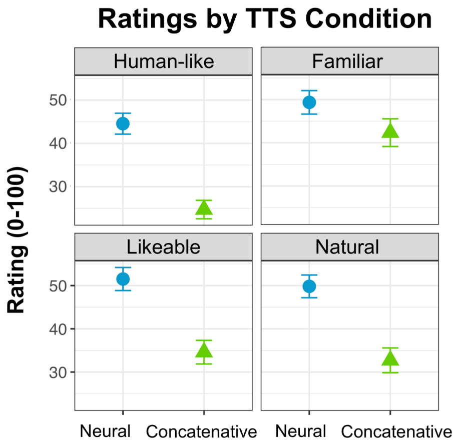

Figure 2: *(Experiment 2) Mean ratings for TTS type: neural (blue circle) vs. concatenative (green triangle). Error bars = standard error.*

3. Experiment 2: Language Attitudes

3.1. Methods

Following Experiment 1, the same participants completed a ratings study, where they heard a single sentence (“The girl knows about the swamp.”) produced by each of the 4 speakers in the 2 TTS conditions (8 voices in total) and provided 4 ratings of the voice using a sliding scale (0-100): (1) How machine-like/human-like, (2) How unfamiliar/familiar?, (3) How eerie/natural? and (4) How unlikeable/likeable? Order of neural and concatenative TTS voices was blocked, so that the TTS for the same speaker was not presented sequentially. The ratings task was also blocked by question: Participants provided a rating for each of the 8 voices for a given dimension (e.g., ‘human-likeness’).

3.2. Analysis & Results

Participants' ratings of the voices were analyzed using separate linear mixed effects models with the *lme4* R package [35]. Fixed effects included TTS Condition; random effects included by-Listener and by-Speaker random intercepts.

As seen in Figure 2, all models showed a similar main effect of TTS Condition: listeners rated the neural TTS voices as more human-like [β=9.92, t=8.6, p<0.001], likeable [β=8.5, t=6.4, p<0.001], natural [β=8.5, t=6.3, p<0.001], and familiar [β=3.50, t=2.0, p<0.05] than the concatenative TTS.

4. Relating Intelligibility and Ratings

To test whether there was a relationship between an individual participant's rating for a given voice (e.g., 'Salli', neural) and their word identification accuracy for that voice, we conducted a post-hoc analysis. We modeled word identification accuracy in separate mixed effects logistic regression models for the 4 ratings (familiar, human-like, natural, and likeable), with the fixed effect of TTS Condition, Rating Score (continuous, z-scored within speaker/rating), their interaction, and by-Subject and by-Speaker random intercepts.

All four models showed no main effect of ratings on accuracy. However, two models revealed significant interactions: word identification accuracy was higher for neural TTS when they were rated as being more human-like [β=0.09, t=2.9, p<0.01] or more natural [β=0.06, t=2.0, p<0.05]. There was no effect of familiarity [β=-0.05, t=1.7, p=0.10] or likeability [β=0.05, t=1.8, p=0.08] by TTS.

5. Discussion

The present study investigated whether the type of TTS synthesis method (concatenative or neural) results in different listener perception patterns. In Experiment 1 (speech-in-noise), we found that neural TTS resulted in overall reduced intelligibility, relative to the concatenative TTS method. This result is in line with prior research indicating that more casual, connected speech results in more difficulty for listeners in identifying the linguistic message in human-human interaction [19], suggesting that it extends to synthesized voices. For one, this finding suggests that neural TTS, while increasingly naturalistic, may actually reduce listeners' ability to understand speech from a modern voice-AI system, if it's being used in the presence of competing noise (e.g., a fan, multiple background talkers). At the same time, this finding counters prior work where increased coarticulation has been shown to improve speech-in-noise perception for TTS voices [20]. While we also used TTS voices in the present study, the synthesis method greatly differed (here, concatenative and neural TTS; formant-based synthesis in [20]). Future work is needed to test what types of coarticulation might be advantageous in more recent TTS methods (e.g., neural TTS). Additionally, this reduction in accuracy for neural TTS was not further modulated by signal-to-noise ratio (SNR) or semantic predictability; in line with prior work [22]–[25], these factors independently reduced accuracy (lower for low predictability; lower at -6 dB SNR) and were additive: accuracy was *further* reduced at low SNR and low predictability.

Meanwhile, in Experiment 2 (social ratings), we observed differences in listeners' ratings of concatenative versus neural TTS for four social attributes: listeners rated neural TTS as more human-like, natural, and familiar, and likeable than concatenative TTS, consistent with prior work (e.g., [2]).

Finally, we linked the data from Experiments 1 & 2. We found individual variation of ratings was linked to word intelligibility. A given listener's ratings of how human-like or natural they found a neural TTS voice correlated with their accuracy in identifying words in that voice: voices that were rated as sounding more natural and more human-like showed *less* of a decrease in intelligibility than voices that were rated as less natural and less human-like. One possible explanation is that a listener who rated the voice as sounding less human-like may assume they would not be able to understand the TTS. This is in line with work on stereotyping of human speakers: where listeners show reduced accuracy in speech-in-noise tasks based on top-down expectations (e.g., [38]). At the same time, listeners might assume that the more 'human-like' TTS voices should also produce the clear speech adaptations that *real* humans produce in more challenging listening conditions (e.g., a lower SNR [36], low semantic predictability [25]), i.e., hyper-speech to the assumed benefit of their listener (cf. H&H Theory: [37]); that these adaptations do *not* occur might be one reason for the lowered accuracy for neural TTS overall (but note that we did not observe any interactions between TTS type and SNR / semantic predictability in the present study). Future work varying the voice qualities across different listening conditions, as well as measuring individual differences (e.g., in computer personification) can tease apart this possible contribution.

There are many other open questions, which can serve as areas for future research. For one, the present study does not include a pre-test to measure listeners' a priori expectations for TTS voices; for example, ratings of the voices (e.g., human-likeness) might have been influenced by the listeners' difficulty hearing that voice in the speech-in-noise study. Additionally, the relative contribution of speakers' experience with voice-AI systems may be a factor in how well they can perceive TTS sentences in noise; in the present study, nearly all listeners (26/28) had prior experience using voice-AI. How this experience might interact with listeners' expectations is also an area to be explored. Moreover, an in-depth investigation of coarticulation patterns would be insightful to quantify *why* neural TTS voices sound more human-like, as well as what coarticulatory adjustments may improve intelligibility.

Finally, this study has implications for voice user interface design. For one, it is noteworthy that the more advanced and realistic TTS method results in less intelligible speech in adverse listening conditions. Further work exploring this effect across different types of background noise (e.g., 1-talker, multitalker babble) and across listeners (e.g., older individuals, hearing-impaired individuals, individuals with autism) can be insightful for tailoring TTS across individuals and communicative scenarios. Additionally, our work suggests having a user choose the voice that sounds most human-like and natural to them may aid intelligibility across listening conditions, even if this speech synthesis method is less intelligible in adverse listening conditions overall.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC and by an Amazon Faculty Research Award to GZ. Thank you to Yulan Liu and Roland Maas for their feedback on the paper.

7. References

[1] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the hurricane challenge.," in *Interspeech*, 2013, pp. 3552–3556.

[2] A. Van Den Oord *et al.*, "WaveNet: A generative model for raw audio.," in *SSW*, 2016, p. 125.

[3] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: A discussion and an evaluation," 2019.

[4] T. Merritt *et al.*, "Comprehensive Evaluation of Statistical Speech Waveform Synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2018, pp. 325–331, doi: 10.1109/SLT.2018.8639556.

[5] S. Ronanki, "Prosody generation for text-to-speech synthesis," Dissertation, University of Edinburgh, 2019.

[6] E. Farnetani and D. Recasens, "Coarticulation and connected speech processes," *The handbook of phonetic sciences*, pp. 371–404, 1997.

[7] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech processing in the auditory system*, Springer, 2004, pp. 231–308.

[8] M. Fallon, S. E. Trehub, and B. A. Schneider, "Children's perception of speech in multitalker babble," *The Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 3023–3029, 2000.

[9] M. N. Ruscetta, E. M. Arjmand, and S. R. Pratt, "Speech recognition abilities in noise for children with severe-to-profound unilateral hearing impairment," *International Journal of Pediatric Otorhinolaryngology*, vol. 69, no. 6, pp. 771–779, 2005.

[10] P. E. Souza and C. W. Turner, "Masking of speech in young and elderly listeners with hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 3, pp. 655–661, 1994.

[11] M. L. G. Lecumberri and M. Cooke, "Effect of masker type on native and non-native consonant perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2445–2454, Mar. 2006, doi: 10.1121/1.2180210.

[12] O. Simantiraki, M. Cooke, and S. King, "Impact of Different Speech Types on Listening Effort.," in *Interspeech*, 2018, pp. 2267–2271.

[13] R. C. Gilbert, B. Chandrasekaran, and R. Smiljanic, "Recognition memory in noise for speech of varying intelligibility," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 389–399, 2014.

[14] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 1, pp. 96–103, 1985.

[15] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *The Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 272–284, 2002.

[16] 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, 1994.

[17] A. R. Bradlow, N. Kraus, and E. Hayes, "Speaking clearly for children with learning disabilities," *Journal of Speech, Language, and Hearing Research*, 2003.

[18] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," 2019.

[19] L. Shockey, "Phonetic and phonological properties of connected speech," The Ohio State University, 1973.

[20] S. Hawkins and A. Slater, "Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech," 1994.

[21] 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, 2013.

[22] C. G. Clopper, "Effects of dialect variation on the semantic predictability benefit," *Language and Cognitive Processes*, vol. 27, no. 7–8, pp. 1002–1020, Sep. 2012, doi: 10.1080/01690965.2011.558779.

[23] S. Kennedy and P. Trofimovich, "Intelligibility, Comprehensibility, and Accentedness of L2 Speech: The Role of Listener Experience and Semantic Context," *Canadian Modern Language Review*, Mar. 2008, doi: 10.3138/cmlr.64.3.459.

[24] S. V. van der Feest, C. P. Blanco, and R. Smiljanic, "Influence of speaking style adaptations and semantic context on the time course of word recognition in quiet and in noise," *Journal of Phonetics*, vol. 73, pp. 158–177, 2019.

[25] C. G. Clopper and J. B. Pierrehumbert, "Effects of semantic predictability and regional dialect on vowel space reduction," *J Acoust Soc Am*, vol. 124, no. 3, pp. 1682–1688, Sep. 2008, doi: 10.1121/1.2953322.

[26] L. E. Humes, B. U. Watson, L. A. Christensen, C. G. Cokely, D. C. Halling, and L. Lee, "Factors associated with individual differences in clinical measures of speech recognition among the elderly," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 2, pp. 465–474, 1994.

[27] J. I. Alcántara, E. J. Weisblatt, B. C. Moore, and P. F. Bolton, "Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome," *Journal of Child Psychology and Psychiatry*, vol. 45, no. 6, pp. 1107–1114, 2004.

[28] K. J. Van Engen, "Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble," *Speech Communication*, vol. 52, no. 11, pp. 943–953, Nov. 2010, doi: 10.1016/j.specom.2010.05.002.

[29] S. V. Berg, A. Panorska, D. Uken, and F. Qeadan, "DECtalk^{TM} and VeriVox^{TM}: Intelligibility, Likeability, and Rate Preference Differences for Four Listener Groups," *Augmentative and Alternative Communication*, vol. 25, no. 1, pp. 7–18, Jan. 2009, doi: 10.1080/07434610902728531.

[30] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1337–1351, 1977.

[31] H. Quené and L. E. Van Delft, "Non-native durational patterns decrease speech intelligibility," *Speech Communication*, vol. 52, no. 11–12, pp. 911–918, 2010.

[32] M. Winn, *Make speech-shaped noise*. 2019.

[33] D. McCloy, *Mix speech with noise*. 2015.

[34] J. Reilly, V. Troiani, M. Grossman, and R. Wingfield, "An introduction to hearing loss and screening procedures for behavioral research," *Behavior Research Methods*, vol. 39, no. 3, pp. 667–672, Aug. 2007, doi: 10.3758/BF03193038.

[35] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, Oct. 2015, doi: 10.18637/jss.v067.i01.

[36] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech communication*, vol. 20, no. 1–2, pp. 13–22, 1996.

[37] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*, Springer, 1990, pp. 403–439.

[38] M. Babel and J. Russell, "Expectations and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2823–2833, 2015.