Journal of
Otorhinolaryngology, Hearing and Balance Medicine

MDPI

Article

The Intelligibility Benefits of Modern Computer-Synthesized Speech for Normal-Hearing and Hearing-Impaired Listeners in Non-Ideal Listening Conditions

Yizhen Ma [1] and Yan Tang [2,3,*]

[1] Department of Linguistics, University of Rochester, Rochester, NY 14627, USA; yma60@ur.rochester.edu
[2] Department of Linguistics, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[3] Beckman Institute for Advanced Science and Technology, Urbana, IL 61801, USA
* Correspondence: yty@illinois.edu; Tel.: +1-217-244-3019

Abstract: Speech intelligibility is a concern for public health, especially in non-ideal listening conditions where listeners often listen to the target speech in the presence of background noise. With advances in technology, synthetic speech has been increasingly used in lieu of actual human voices in human–machine interfaces, such as public announcement systems, answering machines, virtual personal assistants, and GPS, to interact with users. However, previous studies showed that speech generated by computer speech synthesizers was often intrinsically less natural and intelligible than natural speech produced by human speakers. In terms of noise, listening to synthetic speech is challenging for listeners with normal hearing (NH), not to mention for hearing-impaired (HI) listeners. Recent developments in speech synthesis have significantly improved the naturalness of synthetic speech. In this study, the intelligibility of speech generated by commercial synthesizers from Google, Amazon, and Microsoft was evaluated by both NH and HI listeners in different noise conditions. Compared to a natural female voice as the baseline, listeners' listening performance suggested that some of the synthetic speech was significantly more intelligible even at rather adverse listening conditions for the NH cohort. Further acoustical analyses revealed that elongated vowel sounds and reduced spectral tilt were primarily responsible for improved intelligibility for NH, but not for HI due to their impairment at high frequencies and possible cognitive decline associated with aging.

Keywords: intelligibility; synthetic speech; hearing-impaired; noise; intelligibility model; Mandarin Chinese

Citation: Ma, Y., Tang, Y. The Intelligibility Benefits of Modern Computer-Synthesized Speech for Normal-Hearing and Hearing-Impaired Listeners in Non-Ideal Listening Conditions. *J. Otorhinolaryngol. Hear. Balance Med.* **2024**, *5*, 5. https://doi.org/10.3390/ohbm5010005

Academic Editors: Toshihisa Murofushi and Yu Sun

Received: 22 February 2024
Revised: 2 April 2024
Accepted: 11 April 2024
Published: 18 April 2024

1. Introduction

Synthetic speech generated by computers has been increasingly used in lieu of actual human voices in human–machine interfaces. Nearly all digital assistants, such as Amazon Alexa and Apple Siri, on contemporary mobile devices, personal computers, and consumer electronics interact with users using synthetic voices. GPS navigators also use synthetic voices to guide their users. However, the information exchange between humans and machines often takes place in non-ideal (e.g., noisy and reverberant) listening conditions. For example, drivers often hear the instructions given by the GPS mixed with surrounding sounds, such as engine, road, and wind noises. Therefore, speech intelligibility in similar scenarios is crucial, especially in safety-critical environments. Investigating the acoustic traits that indicate speech intelligibility has continued for decades; measuring and improving the intelligibility of synthetic speech has been heavily studied since the emergence of speech synthesizers [1,2]. In the past, it was found that synthetic speech was less intelligible than natural speech in quiet conditions [2] and was even less so in non-ideal listening conditions [3,4]. As a public health concern, the intelligibility issue of synthetic voices in everyday life, especially for the hard-of-hearing population, has drawn considerable attention.

Enormous efforts have been devoted to designing novel algorithms for speech synthesis. The earliest speech synthesizers were developed from rule-based models that filter a source signal to resemble speech-to-data-based models that concatenate segments from recorded human voices [5,6]. However, speech generated by these synthesizers is highly unnatural, especially in the prosody domain. Statistical parametric synthesis (SPS) was proposed to address these problems [7]. SPS and subsequently developed neural models directly synthesize speech waveforms instead of concatenating units from natural speech. A typical SPS pipeline includes three components: a text analyzer that extracts linguistic features from the text, an acoustic model that translates linguistic features to acoustic features, and a vocoder that outputs a waveform based on the acoustic features. The neural synthesizers that came later attempted to simplify the above pipeline and strove toward directly producing waveforms from the text. Synthetic voices generated by state-of-the-art neural synthesizers have demonstrated significantly better quality and naturalness compared to traditional SPS speech [7,8].

Concerning speech intelligibility in general, there have been many studies and investigations into which acoustic features are crucial to it and why. For clear speech, for instance, the speaking rate and vowel space are negatively and positively correlated with intelligibility, respectively [9,10]. Moreover, the speaking rate also has an impact on vowel space. Typically, as the speaking rate increases, speech intelligibility and vowel space decrease [11]. The harmonics-to-noise ratio, a measure of voice quality, is also a strong predictor of speech intelligibility [12–14].

Many studies have also been conducted on intelligibility in noise for both natural speech and synthetic speech. When speaking in noise, humans spontaneously make articulatory changes to produce speech adapted to the listening environment, known as "Lombard speech" [15]. Lombard speech has been extensively studied, and it is often more intelligible than ordinary speech produced in quiet. Compared to ordinary speech, Lombard speech typically has increased intensity and fundamental frequency (F0), prolonged vowel duration, and an upward shift of energy towards the middle or higher frequencies [16,17]. These acoustic changes are understood to help speech sounds better overcome the masking effect of noise. In the early 1980s, the intelligibility of speech generated by two rule-based synthesizers and that of a natural voice in white noise were compared. Although the intelligibility of some phonemes was statistically similar for all three voices, the advantage of natural speech was evident [2]. A more recent study evaluated the intelligibility of algorithmically-modified natural speech and SPS speech in a series of noise conditions. Some of the synthesizers were particularly adapted to produce synthetic speech with the acoustics properties of Lombard speech. Results showed that natural speeches with spectral modification and reduced dynamic range could outperform Lombard speech in noise. Synthetic speech was, however, always less intelligible than natural speech in noise [4]. Further investigation showed that enhancing important regions (e.g., formants) in the spectral envelope and using Lombard-adapted duration and excitation for synthesis could improve the intelligibility of synthetic speech [18]. However, it was also noted in [18] that the improvement of intelligibility in noise due to acoustic modifications came at the cost of naturalness and sound quality of the speaker's voice when perceived in quiet. Nevertheless, speech with better intelligibility tends to be rated of higher quality by listeners in noise [19].

Until recently, the intelligibility of synthetic voices was mostly assessed on normal hearing (NH) listeners. A handful of studies investigated the intelligibility of synthetic speech to hearing-impaired (HI) listeners. In the early 1990s, one of the earliest studies of this kind suggested that the intelligibility of synthetic speech varied, but some synthesizers could reach similar intelligibility of natural speech for HI listeners in *quiet* environments [20,21]. A more recent study [22] suggested that the degree of hearing loss at high frequencies (>8 kHz) for listeners was negatively correlated with synthetic speech's intelligibility. There is also evidence that, for listeners with cochlear implants, synthetic speech with a slow speaking rate was significantly more intelligible than that with a normal rate, which was

understandably more intelligible than that with a fast rate [23]. After synthesis techniques rapidly advanced for over a decade, the intelligibility of synthetic speech by the most recent synthesizers has hardly been reported. It is particularly meaningful to know whether the improvement in the quality and naturalness of synthetic speech seen in recent years could lead to intelligibility gain in non-ideal conditions for both NH and HI listeners.

This study aimed to systematically evaluate the intelligibility of synthetic voices generated by modern commercial speech synthesizers in a series of noise conditions for NH and HI listeners. Six acoustic properties, including speaking rate, F0, harmonic-to-noise ratio, vowel-to-consonant duration ratio, spectral tilt, and vowel space, were measured for the voices and further analyzed to discover and understand how they affected the intelligibility of synthetic speech for the two cohorts of listeners. The main findings and implications were further discussed.

2. Speech Stimuli and Conditions

2.1. Speech Synthesizers

Three state-of-the-art commercial neural speech synthesizers were chosen to generate synthetic voices to be evaluated. The Amazon Polly neural text-to-speech (TTS) system first converts sequences of phonemes to spectrograms, chooses the spectrograms in which the energy is distributed more in human-sensitive frequencies, and then translates these spectrograms to sound signals by a neural vocoder [24]. The Azure neural TTS voices are synthesized by a modification of the FastSpeech2 model, in which the acoustic model produces spectrograms based on pitch and duration predictors, and the vocoder converts the spectrograms to signals [25,26]. The Google Cloud TTS system implements the WaveNet2 model and directly produces speech one signal at a time by a convolutional neural network [27,28].

Speech was generated in both male and female voices, but the male voices were chosen for the Google and Microsoft Azure synthesizers because they were predicted by an intelligibility model (see Section 2.4) to be more intelligible than their female counterparts in the same noise environments. Amazon Polly only supported a female voice in Mandarin. For comparison, a female voice was also recorded to balance the sex distribution in the four voices, as further detailed in Section 2.2. Table 1 lists information about these four voices.

Table 1. Synthesizers and voices used to generate speech for evaluation.

Synthesizer	Voice Model	Sex	Is Synthetic?	Referred to as
Amazon AWS Polly	"Zhiyu"	Female	Yes	Amazon
Microsoft Azure	"YunxiNeural"	Male	Yes	Microsoft
Google TTS	"cmn-CN-Wavenet-B"	Male	Yes	Google
Natural Voice	-	Female	No	baseline

2.2. Speech Materials

The corpus used for generating the synthetic speech was the Mandarin Speech Perception (MSP) test materials [29]. This corpus contains a total of 100 sentences, and each sentence consists of seven monosyllabic characters. The sentences have sensible meanings, and the words used are common in everyday life. Hence, they have moderate predictability. The entire corpus is phonemically balanced. The distribution of phonemes and tones in this corpus is comparable to that of commonly used Mandarin Chinese words [29].

Each of the three commercial speech synthesizers was used to generate the synthetic version of the 100 MSP sentences, resulting in three sets of MSP sentences. All the speech signals were synthesized and saved in mono WAV format, with a sample size of 16 bits and a sampling rate of 16 kHz. To set a baseline using natural human voice, a female native Chinese speaker recorded the 100 sentences in a sound-treated audio booth using an AKG C520 condenser microphone. During recording, the microphone was attached to the speaker's head, and the head of the microphone was placed 2 cm away from the

corner of the speaker's mouth. The audio signals captured by the microphone were further preamplified by an RME Fireface UCX II audio interface before being saved in WAV format with a sampling rate of 44.1 kHz. When generating stimuli for evaluation, the recordings were downsampled to 16 kHz; the intensity in root-mean-square of all WAV files, including the synthetic sounds, was normalized to 0.01 to avoid clipping during writing the files to the hard drive.

2.3. Noise Maskers and Stimuli

Two types of noise were chosen to mix with speech utterances: speech-shaped noise (SSN) and speech-modulated noise (SMN). The SSN was generated by filtering white noise using the average vocal tract effect of 120 randomly selected utterances from the four sets of sounds. A 100th-order linear predictive coding was used to estimate the vocal tract effect. The SSN consequently shared the long-term average spectrum of the combined corpora, which maximized the masking effect on the target speech in the frequency domain. To generate SMN, the speech envelope was first extracted and smoothed by convolving a 10-min utterance with a 7.2-ms-long pulse train. Then, the SMN became the SSN scaled by the speech envelope. In the time domain, while SSN is temporally stationary, the waveform of SMN resembles that of an actual speech signal with temporal modulations, as illustrated in the left panel of Figure 1. Despite their identical spectra as shown in the right panel of Figure 1, SSN and SMN are known to have different masking effects on the target speech even under the same speech-to-noise ratios (SNRs).

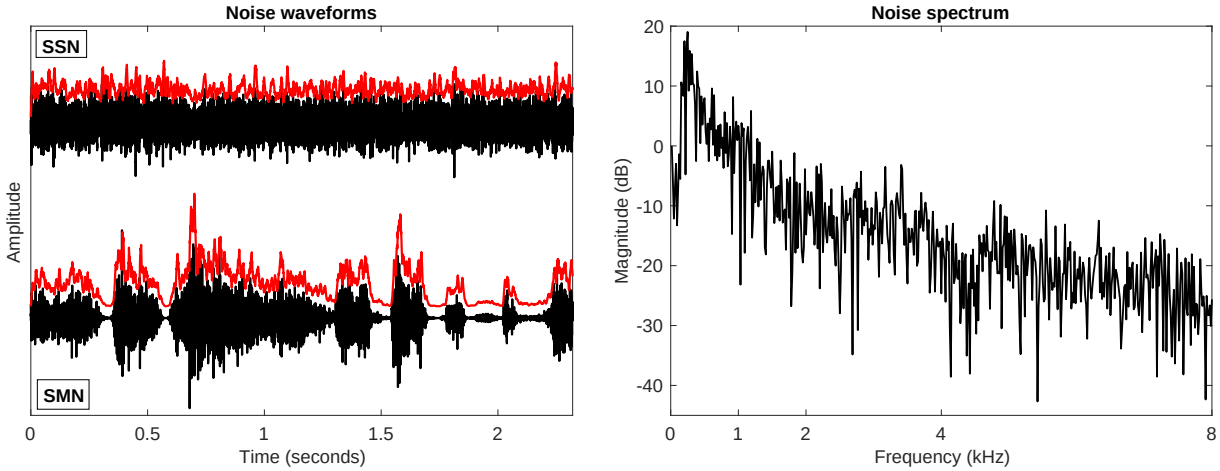

Figure 1. Examples of SSN (**top left**) and SMN waveforms (**bottom left**) and their spectra (**right**). The temporal envelopes of the noise maskers are also displayed in red, offset from the waveforms for better illustration.

Target speech signals were mixed with each noise masker at three SNRs as stimuli. As the two maskers vary in their masking effect as mentioned above, the chosen SNRs were −9, −4, and 1 dB for SSN and −12, −7, and −2 dB for SMN, determined empirically and referred to as "low", "mid", and "high" hereafter. The chosen three SNRs were expected to lead to listener intelligibility of approximately 25%, 50%, and 75%, respectively, in each masker [4,30].

2.4. Model Predictions

The intelligibility of the voices in the noise conditions was estimated using an object intelligibility model—Distortion-Weighted Glimpse Proportion (DWGP, [31])—as the first approximation. Target speech and its masking counterpart are taken as the inputs for DWGP to generate auditory spectro-temporal (S-T) representations for the signals. Auditory analyses are performed by the model to determine the number of S-T regions in the speech signal with a local SNR above a 3-dB threshold, known as "glimpses", at 34 frequencies that simulate auditory filtering. For each frequency band, the similarity between the clean

target speech and speech-plus-noise mixture is also measured. The final predictive index is a logarithmically compressed linear combination of the glimpse quantities weighted by the similarity scores in their frequency bands. From 0 to 1, greater DWGP scores predict better intelligibility. Figure 2 displays the DWGP score of the estimated intelligibility as a function of SNR for the four voices in the chosen noise maskers. In Figure 2, each data point is an average of DWGP scores from the 100 MSP sentences in the same condition. The model predicted that "Amazon" performed consistently better than the other three voices at all SNRs tested in both SSN and SMN. "Google" and "Microsoft" were predicted to have similar intelligibility; "baseline", i.e., the natural voice, was the least intelligible.

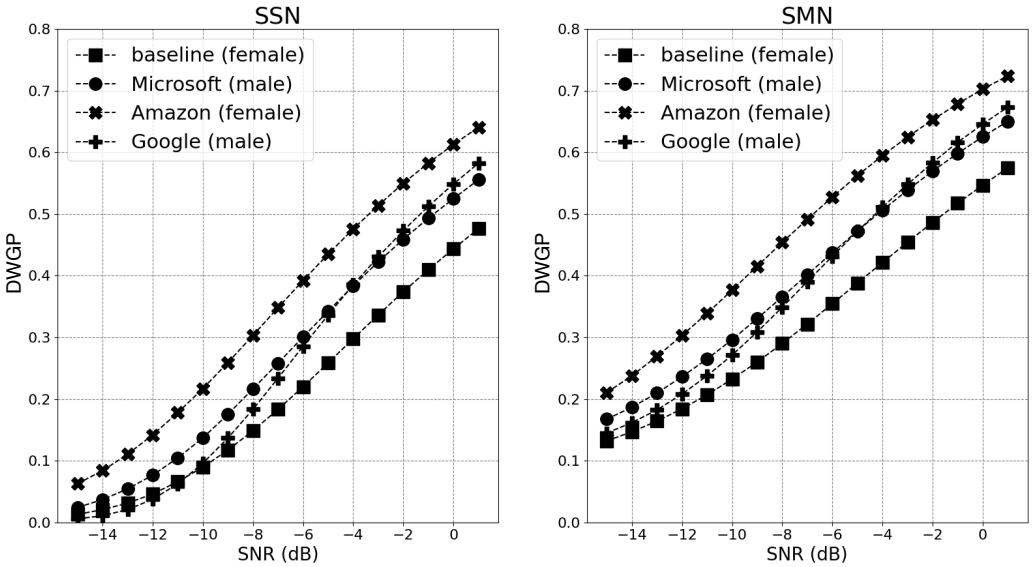

Figure 2. Predicted intelligibility scores from DWGP for all four voices in SSN (**left**) and SMN (**right**).

3. Listening Experiments

3.1. Participants

Ten native Mandarin Chinese speakers (6 females, 4 males, average age = 23.5 years with s.d. = 2.3 years) with self-reported NH were recruited. All the participants were undergraduate or graduate students at the University of Illinois at Urbana-Champaign. The recruitment and testing methods were approved by the Institutional Review Board of the university. All the participants were compensated financially or via course credit for their time. A post-experiment hearing screening was conducted on all the participants using standard clinical audiometric procedures with an audiometer (Interacoustics AS608e). Their hearing levels (HL) were measured as pure-tone average (PTA6) over the six octave frequencies (250, 500, 1000, 2000, 4000, and 8000 Hz). The results showed that one female participant had subtle hearing loss with a PTA6 above 20 dB HL, leading to her data being excluded from further analysis.

Eight native Mandarin Chinese speakers (5 females, 3 males, average age = 57.6 years with s.d. = 6.4 years) with symmetrical sensorineural hearing loss also participated in this experiment in Jiangsu, China. Of the eight participants, four were daily hearing aid users. Their HLs were measured using the same equipment and procedure as for the NH cohort. The PTA6 across the eight symmetrical cases was 36 dB HL for both the left and right ears, with an interaural difference below 3 dB. The mean HLs in PTA6 for the HI cohort are presented in Figure 3.

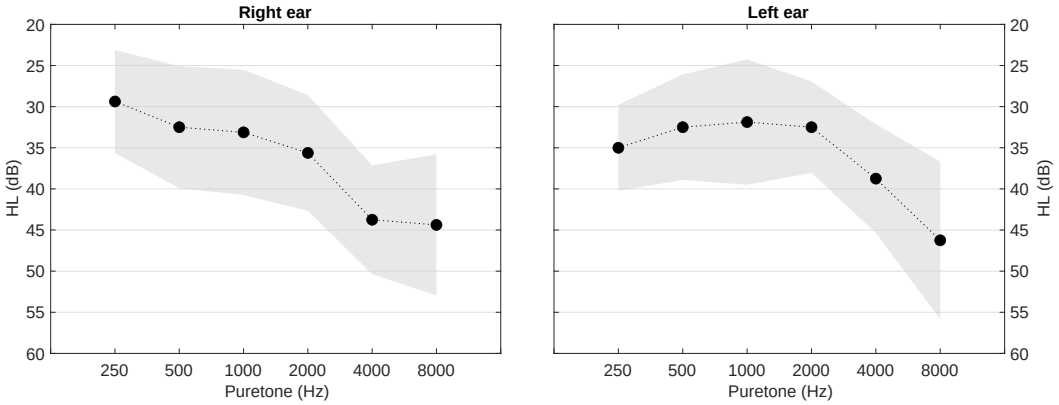

Figure 3. Average two-ear HLs (dB) of the eight HI listeners with symmetrical hearing loss at six octave frequencies. The solid circles and shades show the mean levels and the corresponding 95% confidence intervals, respectively.

3.2. Procedure

A pool of 2400 stimuli (4 voices × 2 noise maskers × 3 SNRs × 100 sentences) was created. In each stimulus, the clean speech was buffered with 300-ms silences at the beginning and the end before mixing with the noise. This allowed listeners' hearing to adapt to the background noise and reduce the effect of non-simultaneous masking [32]. The actual SNR was calculated only from where the target speech was present.

Given the 4 voices, 2 noise maskers, and 3 SNRs, this design led to a total of 24 conditions. In each condition, a listener heard 4 random sentences, resulting in 96 unique sentences screened throughout the test. As some sentences may be more intrinsically intelligible than others, sentences were randomly drawn *without* placement from the pool for each condition, in order to minimize such an effect. The sampling was also validated to make sure none of the 96 sentences was heard twice by the same listener in different conditions. During the test, sentences were blocked by noise-SNR conditions. Within each block, the order of the sentences was randomized; the order of the noise-SNR blocks was also randomized.

For NH listeners, the experiments took place in the same sound-treated audio booth where the natural voice was recorded, as described in Section 2.2. The stimuli were presented to listeners over a pair of Sennheiser HD 660S open-back headphones, preamplified by the same audio interface used in recording. The presentation level of speech was calibrated and fixed to approximately 69 dB SPL—the normal conversational level [33]; the noise level was then adjusted to meet the desired SNRs. Stimulus playback was administered by a computer program. After a stimulus was played, participants were instructed to type what they heard from the sentence in Chinese characters using a computer keyboard. Listeners were not allowed to replay any stimulus. Responses were recorded and saved automatically as the experiment progressed. Participants were able to complete the test within 40 min.

Although the HI listeners were recruited at a different site, the experiments were conducted in an audio booth that had similar acoustic specifications as the one for NH listeners. All the experiment materials, design, setup, and equipment strictly followed those for NH listeners, except for the following. All HI listeners wearing hearing aids were instructed to remove their hearing devices and tested with bare ears. Instead of using a fixed presentation level for target speech, all participants in this group were allowed to adjust the speech intensity to a comfortable level (average level = 76.1 dB SPL with s.d. = 5.7 dB) before the experiment started; they could not adjust the level during the test. As the participants in this group had a significantly greater age than the NH group, and some participants were unable to use computers; therefore, the participants responded to stimuli by orally repeating what they heard after each sentence, which was recorded simultaneously by a digital audio recorder.

The intelligibility of the four voices was measured as listener character recognition rate (CRR) in percentages. As Chinese characters have no predictable pattern to their corresponding pronunciations, it is possible that a participant will hear the sound correctly but type a character with the same pronunciation that mismatches the meaning. Thus, both responses and keys were converted to Hanyu Pinyin (the phonetic symbols for Chinese characters) before comparing, in order to take homophones into account. The correctness of the lexical tones in Mandarin Chinese was not taken into account, so the syllables with the same phonemes as the answer but different tones were still considered to be correct. Responses of NH listeners were scored automatically using a computer script, while a manual transcribing and scoring procedure had to be used for HI listeners due to their responses being saved in audio.

4. Results

Figures 4 and 5 show the average CRRs of participants for the four voices in different noise conditions for NH and HI listeners. For NH listeners, in SSN, "baseline" appears to be the least intelligible at the low and mid SNRs, with CRRs of 4% and 38%. This is broadly consistent with DWGP predictions in Section 2.4. "Amazon", on the other hand, leads to noticeably higher intelligibility (38%) than "Microsoft" (13%) and "Google" (8%) at low SNR. "Amazon" (85%) is also more intelligible than "Microsoft" (55%) at the mid SNR, but it no longer outperforms "Google" (86%). At the high SNR, all voices lead to CRRs of no lower than 83%, with "Microsoft" being the worst (83%) and "Google" being the best (98%). In SMN, which has a much stronger fluctuation in amplitude than SSN, "Microsoft" is seen to be the least intelligible instead of "baseline" at all SNRs. The advantage of "Amazon" (54%) over other voices ("Google" (47%) and "baseline" (33%)) at the low SNR is less evident than in SSN; the same pattern can also be observed at the mid SNR.

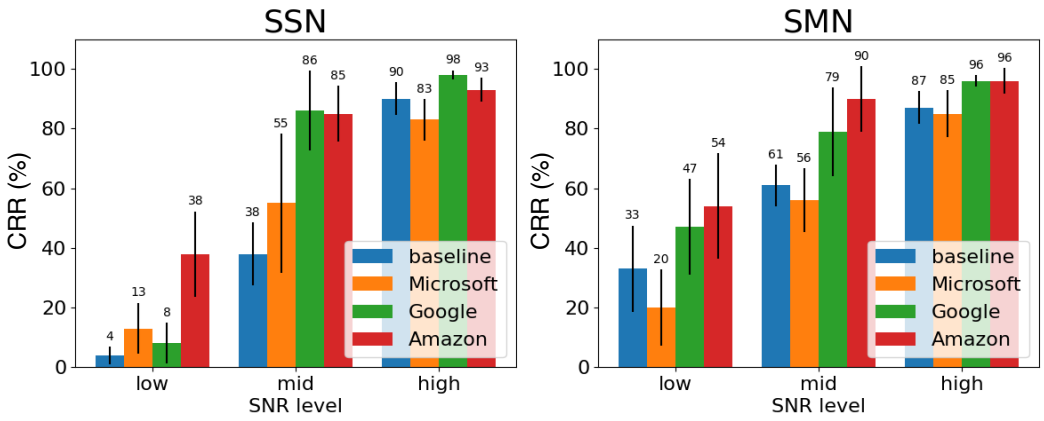

Figure 4. Mean **NH** intelligibility of synthetic speech generated by three commercial synthesizers and natural speech at three SNRs in SSN (**left**) and SMN (**right**). Error bars indicate 95% confidence intervals. Numbers above error bars show the actual CRR readings in percentage for conditions.

For HI listeners, the intelligibility of all voices is unsurprisingly lower than that of their NH counterpart in all conditions, with average decreases of 37.4 percentage points (ppts) in SSN and 38.3 ppts in SMN across SNRs and voices. The variance in performance also tends to be greater in several conditions as indicated by the extended error bars, as exhibited in Figure 5. Interestingly, HI listeners did not benefit from "Amazon" or "Google" as much as their NH counterparts did at the low and mid SNRs in both noise maskers. Compared to "baseline" in SSN, the CRR gain of "Amazon" is merely 2 and −1 ppts for HI listeners at the two lower SNRs, respectively, down from 34 and 47 ppt for the NH cohort in the corresponding conditions. The advantage of synthetic voices over "baseline" further diminishes, with the largest intelligibility gain of only 8 ppts achieved by "Google" at the mid SNR.

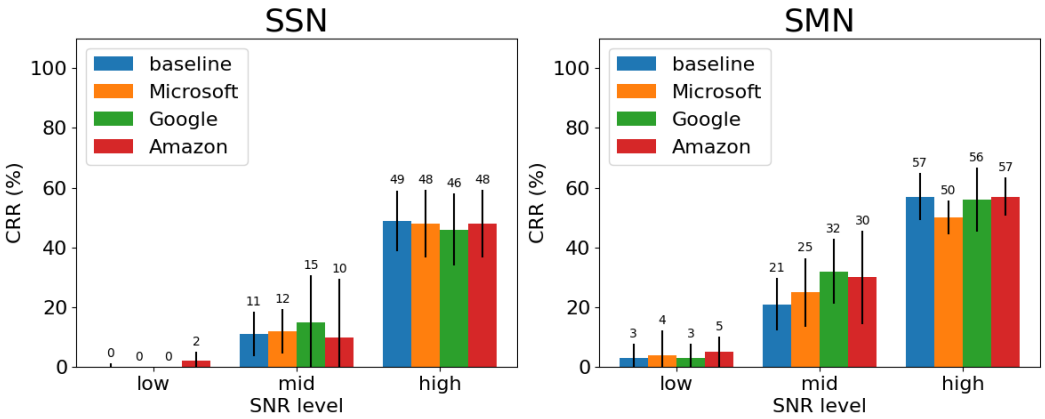

Figure 5. Mean **HI** intelligibility of synthetic speech generated by three commercial synthesizers and natural speech at three SNRs in SSN (**left**) and SMN (**right**). Error bars indicate 95% confidence intervals. Numbers above error bars show the actual CRR readings in percentage for conditions.

Listener performance in CRR of both NH and HI groups was further analyzed together using a multi-level linear mixed effects model (LMM) to account for the within-group and between-group variability. The LMM was implemented using the "lme4" package [34] in R [35], with CRR as the dependent variable and subject variability as the random effect. The four independent variables—group ("NH" and "HI"), noise type ("SSN" and "SMN"), SNR ("low", "mid", and "high") and voice ("Amazon", "Google", "Microsoft", and "baseline")—and their interactions were included as fixed effects. Since all independent variables were categorical, a dummy coding scheme was employed, with group: "NH", noise type: "SSN", SNR: "low", and voice: "baseline" being the reference level in each variable.

A series of linear models was built using a step-wise sequential approach with a forward direction, where the independent variables and their interactions were added incrementally, starting with only the random effect. Table 2 shows the models whose fitting significantly improved over previous models. The results suggested that all independent variables (group, noise, SNR, and voice) had a significant impact on listeners' CRR performance $[\forall \chi^2 \geq 10.399, \forall p < 0.01]$. Most of the two-way and three-way interactions between the independent variables were also significant $[\chi^2 \geq 9.464, p < 0.01]$, and so was their four-way interaction $[\chi^2(50) = 3.213, p < 0.001]$. The final LMM, therefore, included 12 terms, as shown in Table. 2.

Table 2. Step-wise build-up of the LMM. The models were assessed by the Akaike information criterion (AIC), Bayesian information criterion (BIC), log-likelihood (LogLik), and the likelihood ratio with the degrees of freedom ($\chi^2(df)$) and significance level (SL ***: $p < 0.001$; **: $p < 0.01$). Only significant factors and interactions are listed.

Model	AIC	BIC	LogLik	$\chi^2(df)$	SL
rand. intercept	3940.455	3952.489	−1967.228		
group	3903.007	3919.052	−1947.504	39.448 (4)	***
noise	3894.608	3914.664	−1942.304	10.399 (5)	**
SNR	3413.510	3441.589	−1699.755	485.098 (7)	***
voice	3350.052	3390.165	−1665.026	69.458 (10)	***
noise:SNR	3344.588	3392.723	−1660.294	9.464 (12)	**
SNR:voice	3328.338	3400.541	−1646.169	28.250 (18)	***
group:SNR	3268.592	3348.817	−1614.296	63.746 (20)	***
group:voice	3204.020	3296.279	−1579.010	70.572 (23)	***
noise:SNR:group	3172.665	3276.958	−1560.333	37.355 (26)	***
SNR:voice:group	3161.487	3289.848	−1548.744	23.178 (32)	***
group:noise:SNR:voice	3145.784	3346.347	−1522.892	51.704 (50)	***

The significant four-way interaction, i.e., group × noise type × SNR × voice, indicates that the voices varied in intelligibility at different SNRs in the two noise maskers between NH and HI listeners. Contrasts were used to break down this interaction in order to gain better insights into the data. All the significant contrasts and their parameters for this interaction are listed in Table 3 and 4, respectively. Contrast 1 shows that the intelligibility gain from "Amazon" over "baseline" was different at the mid SNR in the two noise maskers for NH and HI listeners. For "Microsoft", while it was more intelligible than "baseline" at the low SNR in SSN, its intelligibility was lower than "baseline" at the same SNR in SMN for NH cohorts; the difference in intelligibility between the two voices was, however, less sizable for HI listeners. A similar relationship was also observed at the mid SNR in the two maskers for "Microsoft", but the overall intelligibility level of this voice was higher than that at the lower SNR, as suggested by Contrasts 2 and 3. Contrast 4 indicates a similar result for "Google" at the mid SNR in the two noise maskers for the two cohorts of listeners as for Contrast 1 for "Amazon".

Table 3. Significant contrasts for the four-way interaction in the LMM model.

Contrast Index	Fixed Effect (Group : Noise Type : SNR : Voice)
1	(NH vs. HI) : (SSN vs. SMN) : (mid) : (baseline vs. Amazon)
2	(NH vs. HI) : (SSN vs. SMN) : (low) : (baseline vs. Microsoft)
3	(NH vs. HI) : (SSN vs. SMN) : (mid) : (baseline vs. Microsoft)
4	(NH vs. HI) : (SSN vs. SMN) : (mid) : (baseline vs. Google)

Table 4. Parameters estimated from the LMM for the contrasts shown in Table 3, including linear coefficient (β), standard error (SE), low and high confidence interval (CI), t value, effect size (r), and significance level (SL ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$).

Contrast Index	β	SE	Low CI	High CI	t Value	r	SL
1	27.728	9.366	9.327	46.129	2.961	0.148	**
2	23.611	9.366	5.210	42.012	2.521	0.126	*
3	23.760	9.366	5.359	42.161	2.537	0.127	*
4	35.665	9.366	17.264	54.066	3.808	0.189	***

Post hoc pairwise comparisons with the Bonferroni adjustment conformed with the visual impressions received from Figures 4 and 5. In the temporarily-stationary noise (i.e., SSN) , at the low and mid SNRs, "Amazon" is substantially more intelligible than "baseline" and "Microsoft" for NH $[\forall p < 0.001]$. "Google" is also more intelligible than "baseline" and "Microsoft" for NH $[\forall p < 0.001]$ at the mid SNR. In the fluctuating noise (i.e., SMN), at the low SNR, "Microsoft" is the least intelligible voice compared to other voices $[\forall p < 0.001]$ for NH, but not at the mid SNR, where its intelligibility is comparable to "baseline" $[p = 1.000]$ but worse than "Google" $[p < 0.01]$ and "Amazon" $[p < 0.001]$. In both noise maskers, all the voices are similarly intelligible at the high SNR $[\forall p = 1.000]$ for NH, where intelligibility reaches the ceiling. For HI, listener performance shows that there are no voices significantly more intelligible than the others at all SNRs $[\forall p = 1.000]$.

5. Acoustic Analyses

The acoustics of the speech signals were analyzed to explain the listeners' different performances when perceiving the four voices. Speaking rate, fundamental frequencies, harmonic-to-noise ratio, vowel–consonant duration ratio, spectral tilt, and vowel space were analyzed for sentences in each voice. The Montreal Forced Aligner [36] was used to generate phoneme- and word-level speech segments from the WAV files, followed by further manual checking and adjustment of the segment boundaries.

– **Speaking rate**. The speaking rate was measured as the count of spoken syllables per second. As each sentence in this corpus consists of seven words (syllables), the speaking rate for a sentence was calculated by dividing the duration of speech in

seconds by seven. Measurements for the four voices are shown in Figure 6a. Pairwise comparisons using the Bonferroni method suggest that the voices significantly differ in speaking rate [$\forall p < 0.001$], with the difference between "baseline" and "Amazon" being the smallest [$p < 0.05$].

- **Fundamental frequency (F0)**. F0 for each sentence was calculated as the mean F0 over all vowels in the sentence. As a tonal language, the F0 for a syllable in Mandarin can fluctuate greatly and take a sharp upturn, downturn, or both. In order to take tone differences into account, F0s were measured at the beginning, middle, and end of each vowel; the F0 for the vowel was calculated as the mean of F0s at these three points. Since every sentence contains seven monosyllabic vowels, the F0s of the vowels were further averaged and computed as the F0 for the sentence. As shown in Figure 6b, all the voices differ from each other in F0 [$\forall p < 0.001$]. Since "baseline" and "Amazon" are female voices, their mean F0s are significantly higher than those of "Microsoft" and "Google" [$\forall p < 0.001$], which are male voices.
- **Harmonic-to-noise ratio (HNR)**. HNR is the ratio of periodic to non-periodic components in a voiced segment of speech in decibels (dB). It is usually used to show the hoarseness and breathiness of a voice; it is also measured as one aspect of voice quality. For each sentence, the HNR was computed as the mean HNR of the seven monosyllabic vowels. The measurements are shown below in Figure 6c. The comparisons suggest that all the voices also differed in HNR [$\forall p < 0.001$]. Interestingly, "baseline"—the natural voice—had significantly better HNR than all synthetic voices.
- **Vowel–consonant (VC) ratio**. The VC ratio for each sentence was measured by dividing the total duration of all vowels by the total duration of all consonants in the sentence. The statistics are shown below in Figure 6d. It appears that the VC ratio varies greatly across sentences, as suggested by the expanded confidence intervals. While "Microsoft", "Google", and "Amazon", the three synthetic voices, see no significant difference in VC ratio [$\forall p > 0.06$], "baseline" has a significantly lower VC ratio than "Amazon" [$p < 0.01$], but not than "Microsoft" [$p = 0.074$] or "Google" [$p = 0.359$].
- **Spectral tilt**. Spectral tilt is one way to quantify how rapidly energy decreases as frequency increases in a speech signal. For analysis in this study, it was measured as the decrease in energy per unit frequency (dB/Hz); that is, the energy difference between the first formant (F1) and the second formant (F2) over the frequency difference between F1 and F2 on the long-term average spectrum of the signal. Therefore, it was the slope of the straight line connecting the two data points of F1 and F2 on the spectrum. The flatter the slope is, the more energy there is at high frequencies relative to that at low frequencies. From "baseline" to "Google" in Figure 6d, spectral tilt increases successively with significant changes [$\forall p < 0.001$]. Consistent with the visual intuition also, the spectral tilt of "Amazon" is evidently the lowest among the four voices [$\forall p < 0.001$].
- **Vowel space**. The vowel spaces and their visualizations for the voices were calculated and drawn using the R package "phonR" [37]. The vowel spaces were calculated by plotting three corner vowels on an F1–F2 chart and finding the smallest polygon that covers the vowels. The three corner vowels are [i], [a], and [u]. These three vowels are frequently supplied as corner vowels in vowel space drawing [11,38] and are all present in Mandarin as monophthongs. These graphs are shown in Figure 7. The vowel space areas (VSAs) of "Amazon", "Microsoft", "baseline", and "Google" are 1,365,952, 1,195,746, 1,096,042, and 981,738, respectively, in descending order. A greater VSA may indicate better speech-motor control and more distinctive articulatory positions for different sounds. To further measure the quality of individual vowels and their overlaps, between- and within-vowel dispersions and their ratios were also measured from the three groups of sound for the four voices. The greater the between-vowel dispersion, the further the different sounds are from each other, hence the less likely they would be confused by the listeners. Meanwhile, a smaller within-

vowel dispersion suggests fewer variations for the same vowel, hence a more robust and consistent pronunciation of the vowel. As illustrated in Figure 7 and further confirmed by Table 5, the three vowels in "baseline" were most distinctive from each other (the largest between-vowel dispersion) and were pronounced most consistently (the smallest within-vowel dispersion), followed by "Amazon". The between-within-vowel dispersion ratio, on the other hand, showed that "Google" had the lowest vowel quality, which is in line with its smallest VSA.

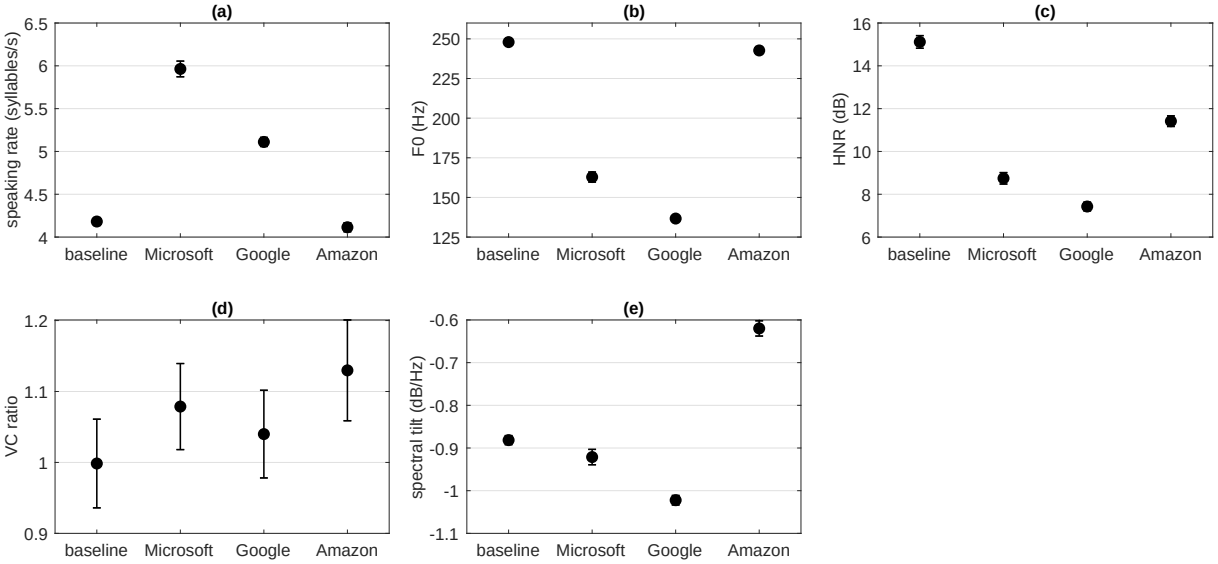

Figure 6. Speaking rate (**a**), F0 (**b**), HNR (**c**), VC ratio (**d**), and spectral tilt (**e**) of the natural and three synthetic voices. Error bars show 95% confidence intervals of the means.

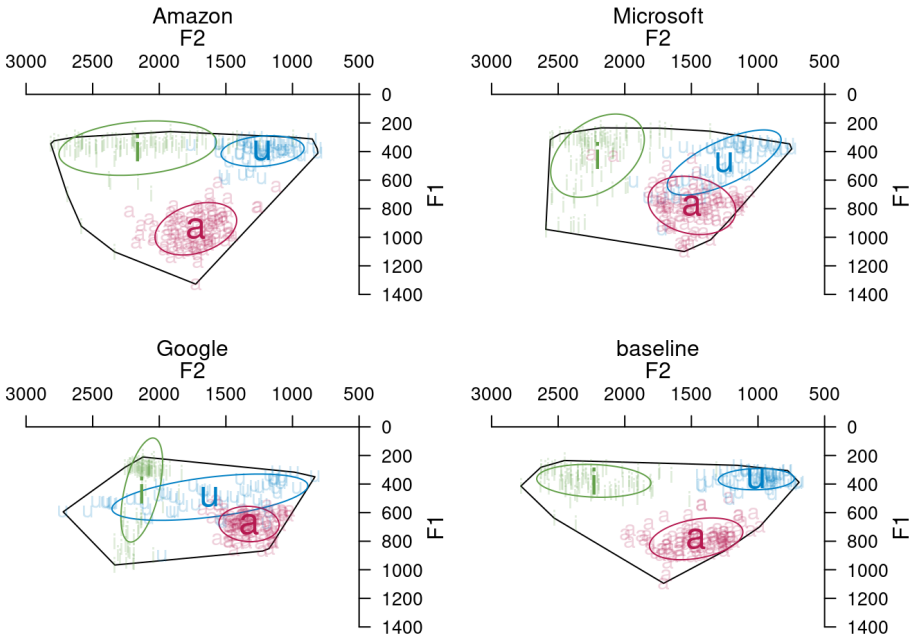

Figure 7. Vowel spaces for the "Amazon", "Microsoft", "Google", and "baseline" voices.

Two general linear models (GLMs) were fitted on the NH and HI data separately to inspect how well speaking rate, F0, HNR, VC ratio, and spectral tilt as independent variables predicted listener performance in the four voices. The measurements of vowel spaces were not included because, instead of at the sentence level, they were derived at the corpus level for each voice. The values of each predictor were centered to have

a mean of zero, further scaled by their standard deviation. For NH listeners, the predictions made by the GLM had a significant correlation with the intelligibility in CRR [$R^2 = 0.130$, $F(210) = 6.299$, $p < 0.001$]. The linear coefficients of the predictors, β, and their SLs displayed in Table 6 suggest that speaking rate, F0, and spectral tilt played significant roles in estimating the intelligibility of the voices, but including HNRs and VC ratios did not significantly improve the model performance. For HI listeners, the chosen predictors together, however, were unable to make predictions significantly correlated with the CRR [$R^2 = 0.010$, $F(186) = 1.385$, $p = 0.232$], although including F0, HNR, and spectral tilt, but not others, still significantly improved the model fitting, as indicated by their β values in Table 7.

Table 5. Vowel quality in the four voices measured for [a], [i], and [u] using between-vowel dispersion ξ_b, within-vowel dispersion ξ_w, and their ratio $\frac{\xi_b}{\xi_w}$.

Voice	ξ_b	ξ_w	$\frac{\xi_b}{\xi_w}$
Amazon	52,729.0	248.7	212.0
Microsoft	49,979.0	250.3	199.7
Google	40,661.0	249.6	162.9
baseline	60,517.0	214.2	282.5

Table 6. Parameters estimated for the linear regression on the **NH** data, including linear coefficient (β), standard error (SE), low and high confidence interval (CI), t value, effect size (r), and significance level (SL ***: $p < 0.001$; *: $p < 0.05$).

Predictor	β	SE	Low CI	High CI	t Value	r	SL
Intercept	62.169	2.083	58.063	66.276	29.843	0.900	***
Speaking rate	−13.577	3.582	−20.640	−6.515	−3.790	0.253	***
F0	−32.653	14.180	−60.605	−4.700	−2.303	0.157	*
HNR	10.318	10.087	−9.566	30.202	1.023	0.070	
VC ratio	2.145	2.226	−2.244	6.533	0.963	0.066	
Spectral tilt	17.466	6.745	4.170	30.762	2.590	0.176	*

Table 7. Parameters estimated for the linear regression on the **HI** data, including linear coefficient (β), standard error (SE), low and high confidence interval (CI), t value, effect size (r), and significance level (SL ***: $p < 0.001$; *: $p < 0.05$).

Predictor	β	SE	Low CI	High CI	t Value	r	SL
Intercept	24.353	1.708	20.983	27.723	14.255	0.723	***
Speaker rate	−2.945	2.974	−8.812	2.921	-0.990	0.072	
F0	−26.429	11.299	−48.720	−4.138	−2.339	0.169	*
HNR	16.836	7.930	1.193	32.480	2.123	0.153	*
VC ratio	1.026	1.850	−2.623	4.676	0.555	0.041	
Spectral tilt	12.000	5.446	1.255	22.745	2.203	0.159	*

6. Discussion

This study examined the intelligibility of speech generated by three commercial synthesizers for NH and HI listeners in different noise conditions. The results were somewhat unexpected but encouraging when compared to listeners' performance in CRR on natural speech produced by a female speaker, as initially predicted by the DWGP model [$R^2 = 0.748$, $p < 0.001$]. Previously, the intelligibility of natural speech was substantially better than that of synthetic speech in noise for NH listeners [2–4]. Listener performance in our experiments, however, showed that the natural voice was the least intelligible among the four voices evaluated in many conditions, especially in the stationary masker (speech-shaped noise or SSN), suggesting that, in the last decade, deep learning-based synthetic speech has improved tremendously in terms of its quality [7] and intelligibility

over that which previously used concatenation, a vocoder, and statistical models. In the temporally-fluctuating noise masker (speech-modulated noise or SMN), the advantage of the synthetic voices over "baseline" was noticeably reduced at the low and mid SNRs. Speech temporal modulation is known to be important to intelligibility. One possibility is that the interference due to the masker's modulation on the target temporal envelope cancelled some benefits from other aspects (see below) and that synthetic modulation could be more susceptible to noise than that of natural speech. "Microsoft" is an example of this, as it appeared to be more intelligible than "baseline" at the low and high SNRs in SSN, but was significantly less intelligible than "baseline" at the low SNR in SMN. Further modulation analysis on the signals may help with understanding this issue better.

For the HI cohort, the correlation between DWGP estimations and listeners' performance [$R^2 = 0.616$, $p < 0.001$] was much lower than that for their NH counterpart. This is not surprising because DWGP makes predictions using parameters, e.g., the outer-ear transfer function and the HL, for NH listeners. Notably, HI listeners' speech understanding did not benefit from any of the state-of-the-art synthesizers as NH did. Nevertheless, it is worth remarking that their intelligibility did not suffer from synthetic speech either compared to natural speech. This shows that modern speech synthesizers are already capable of producing speech sounds that may replace the natural human voice without compromising intelligibility, at least for listeners with normal or even mild hearing loss. Since all hearing aid users were tested without their hearing devices, it is unclear if they could achieve better listening performance with their hearing aids on. Often, hearing devices are fitted to a listener's hearing profile in order to maximize their performance. Therefore, there could also be a possibility that synthetic sounds interact with their hearing devices acoustically, e.g., overly boost or attenuate certain frequencies, leading to the processed sounds having mismatching frequencies for the listener, hence reduced intelligibility.

Further acoustic analyses revealed that speaking rate was one of the significant factors that explained the NH listeners' intelligibility in this study. It was suggested that NH listener performance benefited from a lower speaker rate [$\beta = -13.577$, $p < 0.001$] in general, but this was not the case for HI listeners [$\beta = -2.945$, $p = 0.323$]. This finding is consistent with what was observed in [23] for NH. With a lower speaker rate, listeners may have more time to process and decode the information conveyed by the sounds, especially when perceiving speech in noise, which usually demands higher cognitive loads [39]. This may partly explain the benefit of "Amazon" to listener intelligibility (Figures 4 and 6a). Among the four voices, the lowest speaking rate was approximately 4.12 syllables/s, which is consistent with the normal speaking rate—4.08 syllables/s [40]—for native Mandarin Chinese speakers. It appears that the speaking rates tested in this study were too high for HI listeners to benefit from. Cognitive decline associated with aging, e.g., [41] of the HI group (average age = 57.6 yrs) might be another factor contributing to their reduced acuity in speech perception. F0 alone was not seen to be correlated with listener intelligibility for NH [$R^2 < 0.001$, $p = 0.901$] or HI [$R^2 < 0.001$, $p = 0.880$], but including it in the GLMs improved the model prediction. There is no direct evidence showing that solo F0 changes have a significant impact on intelligibility for NH [42,43] or HI listeners [44]. For Lombard speech, which is often produced when a listener speaks in noise and is understood to be more intelligible than speech produced in quiet, increased F0 was indeed observed together with many other acoustic changes, such as reduced spectral tilt and elongated vowel sounds [45], in relevant studies. Therefore, it is possible that F0's contribution to intelligibility is only manifested in the presence of other prominent acoustic correlates. Meanwhile, it is well-documented that solely reducing spectral tilt can lead to improved speech intelligibility for NH listeners, e.g., [17,46], which is also observed in the current study [$R^2 < 0.034$, $p < 0.01$], especially for "Amazon". In the GLM, it also indicates that intelligibility is proportional to spectral tilt [$\beta = 17.466$, $p < 0.05$]. Despite the significant effect in the GLM model for HI, like F0, spectral tilt was not correlated with intelligibility for HI [$R^2 = 0.002$, $p = 0.551$]. In SSN and SMN, which have a similar spectrum as speech signals, voices with reduced spectral tilt often stand a better chance of escaping

from masking since more energy is located at high frequencies of speech. NH listeners can exploit the unmasking at high frequencies for better intelligibility. As shown in Figure 3, the majority of HI participants in this study had more severe hearing loss at high frequencies, which led to them not being able to take as much of an advantage as NH listeners did.

VSA and vowel quality were not found to have an evident correlation with intelligibility in this study. In general, "Google" was the second most intelligible voice after "Amazon" across all conditions and sometimes appeared to be even better than "Amazon" (Figures 4 and 5). However, the VSA, between-vowel dispersion, and between-within-vowel dispersion ratio of "Google" were the smallest among the four voices. The findings in terms of the relationship between VSA and its pertinent measures and intelligibility are mixed. While expanded VSA was reported in some studies on clear speech, e.g., [10,47], and on Lombard speech compared to normal speech, e.g., [10], reduced VSA was also found in Lombard speech in other studies [48,49]. Therefore, the current findings lean towards the conclusion that VSA and vowel quality do not have a direct impact on the intelligibility of synthetic speech, despite the plausible speculation that little overlap between sounds and consistent pronunciation may reduce confusion between sounds.

This study has a few limitations. First, only six acoustic features were analyzed. There are several other properties that also correlate with intelligibility, including emphasis and prosody [50,51]. Two acoustic measures that can indicate emphasis are F0 contour and maximum intensity of the signals [52], as they give strong cues to listeners of what to expect in the speech. On the same note, the DWGP model predicted intelligibility well, implying that there could be other acoustic aspects accounted for by the DWGP model but not analyzed in this study. For example, no evidence was found to explain the intelligibility ranking for "Google", which led to comparable intelligibility as "Amazon" in several conditions but considerably differed in acoustic measurements analyzed in this study. Second, only a limited number of synthetic voices was evaluated. Including more voices could exhibit different intelligibility and analytical results. Third, the number of participants in both groups was relatively small, which might lead to weak statistical power in the analyses.

7. Conclusions

In this study, the intelligibility of synthetic speech generated by three modern commercial synthesizers was evaluated along with a natural female voice for both NH and HI listeners in noise conditions. For NH listeners, some synthetic voices were significantly more intelligible than natural speech, especially in more adverse conditions, showing that synthetic speech has improved tremendously in both naturalness and intelligibility over the last decade. Subsequent acoustic analyses on the voices revealed that synthetic speech with a slower speaking rate and reduced spectral tilt tends to be more intelligible than others for NH listeners. HI listeners, however, benefited little from those acoustic changes due to their hearing loss at high frequencies and potential cognitive decline with aging. These findings may provide a guideline for future customizable speech synthesis techniques that aim to generate voices according to listeners' hearing profiles.

Author Contributions: All authors were involved in conceptualization, data analysis, and reviewing and editing the manuscript; Y.M. was responsible for conducting the experiments on the NH listeners and composing the original draft; Y.T. sought external assistance to repeat the same experiments on the HI cohort and helped with further improving the writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the University of Illinois Urbana-Champaign (protocol number 22138, approved on 10 August 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study may be available on request from the corresponding author.

Acknowledgments: The authors would like to thank Jian Gong for recruiting and repeating the experiments on the HI participants.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
BIC	Bayesian information criterion
CI	confidence interval
CRR	character recognition rate
DWGP	Distortion-Weighted Glimpse Proportion
F0	fundamental frequency
F1	the first formant
F2	the second formant
GLM	general linear model
HNR	harmonic-to-noise ratio
NH	normal hearing
HI	hearing-impaired
HL	hearing level
LMM	linear mixed-effect model
PTA	pure tone average
SI	speech intelligibility
SL	significance level
SNR	speech-to-noise ratio
SII	Speech Intelligibility Index
SSN	speech-shaped noise
SPS	statistical parametric synthesis
SMN	speech-modulated noise
TTS	text-to-speech
VC	vowel–consonant
VSA	vowel space area

References

1. Fant, C.G.M. *Analysis and Synthesis of Speech Processes*; North-Holland Publishing Comp.: Amsterdam, The Netherlands, 1968; pp. 32–58.
2. Clark, J.E. Intelligibility comparisons for two synthetic and one natural speech source. *J. Phon.* **1983**, *11*, 37–49. [CrossRef]
3. Nixon, C.W.; Anderson, T.R.; Moore, T.J. The Perception of Synthetic Speech in Noise. In *Basic and Applied Aspects of Noise-Induced Hearing Loss*; NATO ASI Series; Salvi, R., Henderson, D., Hamernik, R., Colletti, V., Eds.; Springer, Boston, MA, USA, 2007; Volume 111, pp. 32–58.
4. Cooke, C.; Mayo, C.; Valentini-Botinhao, C.; Stylianou, Y.; Sauert, B.; Tang, Y. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Commun.* **2013**, *55*, 572–585. [CrossRef]
5. Black, A.W.; Zen, H.; Tokuda, K. Statistical Parametric Speech Synthesis. In Proceedings of the ICASSP, Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. IV1229–IV1232.
6. Taylor, P. *Text-to-Speech Synthesis*; Cambridge University Press: Cambridge, UK, 2009.
7. Tan, X.; Qin, T.; Soong, F.; Liu, T.Y. A Survey on Neural Speech Synthesis. 2021. Available online: https://arxiv.org/abs/2106.15561 (accessed on 12 February 2024).
8. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. 2016. Available online: https://arxiv.org/abs/1609.03499 (accessed on 25 January 2024).
9. Bergeson, T.R.; Miller, R.J.; McCune, K. Mothers' Speech to Hearing-Impaired Infants and Children With Cochlear Implants. *Infancy* **2006**, *10*, 221–240. [CrossRef]
10. Tang, P.; Xu Rattanasone, N.; Yuen, I.; Demuth, K. Phonetic enhancement of Mandarin vowels and tones: Infant-directed speech and Lombard speech. *J. Acoust. Soc. Am.* **2017**, *142*, 493–503. [CrossRef] [PubMed]
11. Turner, G.; Tjaden, K.; Weismer, G. The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *J. Speech Lang. Hear. Res.* **1995**, *38*, 1001–1013. [CrossRef] [PubMed]

12. Bradley, J.S.; Reich, R.D.; Norcross, S.G. On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. *J. Acoust. Soc. Am.* **1999**, *106*, 1820–1828. [CrossRef] [PubMed]
13. Freyman, R.L.; Griffin, A.M.; Oxenham, A.J. Intelligibility of whispered speech in stationary and modulated noise maskers. *J. Acoust. Soc. Am.* **2012**, *132*, 2514–2523. [CrossRef] [PubMed]
14. Latham, H.G. The signal-to-noise ratio for speech intelligibility – An auditorium acoustics design index. *Appl. Acoust.* **1979**, *12*, 253–320. [CrossRef]
15. Junqua, J.C. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Commun.* **1996**, *20*, 13–22. [CrossRef]
16. Castellanos, A.; Benedí, J.; Casacuberta, F. An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Commun.* **1996**, *20*, 23–35. [CrossRef]
17. Lu, Y.; Cooke, M. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Commun.* **2009**, *51*, 1253–1262. [CrossRef]
18. Valentini-Botinhao, C.; Yamagishi, J.; King, S. Intelligibility Enhancement of Speech in Noise. In Proceedings of the the Institute of Acoustics, Birmingham, UK, 14–15 October 2014; Volume 36.
19. Tang, Y.; Arnold, C.; Cox, T.J. A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners. *J. Otorhinolaryngol. Hear. Balance Med.* **2017**, *1*, 5. [CrossRef]
20. Kangas, K.; Allen, G. Intelligibility of synthetic speech for normal-hearing and hearing-impaired listeners. *J. Speech Hear. Disord.* **1990**, *55*, 751–755. [CrossRef] [PubMed]
21. Humes, L.E.; Nelson, K.J.; Pisoni, D.B. Recognition of synthetic speech by hearing-impaired elderly listeners. *J. Speech Hear. Res.* **1991**, *34*, 1180–1184. [CrossRef] [PubMed]
22. Wolters, M.; Campbell, P.; DePlacido, C.; Liddell, A.; Owens, D. The Effect of Hearing Loss on the Intelligibility of Synthetic Speech. In Proceedings of the 16th ICPhS, Saarbrücken, Germany, 6–10 August 2007; pp. 673–675.
23. Ji, C.; Galvin, J.J.I.; Xu, A.; Fu, Q.J. Effect of Speaking Rate on Recognition of Synthetic and Natural Speech by Normal-Hearing and Cochlear Implant Listeners. *Ear Hear.* **2013**, *34*, 313–323. [CrossRef] [PubMed]
24. Neural TTS. 2024. Available online: https://docs.aws.amazon.com/polly/latest/dg/NTTS-main.html (accessed on 25 January 2024).
25. Liao, Q.Y.; Li, B.H.; Liu, Y.Q.; Tan, X.; Zhao, S. Introducing the Latest Technology Advancement in Azure Neural TTS: Uni-TTSv3. 2021. Available online: https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/introducing-the-latest-technology-advancement-in-azure-neural/ba-p/2595922 (accessed on 8 February 2024).
26. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. 2022. Available online: https://arxiv.org/abs/2006.04558 (accessed on 8 February 2024).
27. Aharon, D. Introducing Cloud Text-to-Speech Powered by DeepMind WaveNet Technology. 2018. Available online: https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-text-to-speech-powered-by-deepmind-wavenet-technology (accessed on 25 January 2024).
28. WaveNet: A Generative Model for Raw Audio. 2016. Available online: https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio (accessed on 25 January 2024).
29. Fu, Q.J.; Zhu, M.; Wang, X. Development and validation of the Mandarin speech perception test. *J. Acoust. Soc. Am.* **2011**, *129*, EL267–EL273. [CrossRef] [PubMed]
30. Tang, Y.; Cooke, M. Learning static spectral weightings for speech intelligibility enhancement in noise. *Comput. Speech Lang.* **2018**, *49*, 1–16. [CrossRef]
31. Tang, Y.; Cooke, M.; Fazenda, B.M.; Cox, T.J. A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers. *J. Acoust. Soc. Am.* **2016**, *140*, 1858–1870. [CrossRef] [PubMed]
32. Marrufo-Pérez, M.I.; del Pilar Sturla-Carreto, D.; Eustaquio-Martín, A.; Lopez-Poveda, E.A. Adaptation to Noise in Human Speech Recognition Depends on Noise-Level Statistics and Fast Dynamic-Range Compression. *J. Neurosci.* **2020**, *40*, 6613–6623. [CrossRef] [PubMed]
33. *ANSI S3.5-1997*; Methods for the Calculation of the Speech Intelligibility Index. American National Standards Institute, Inc.: Washington, DC, USA, 1997.
34. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
35. R Core Team. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing, Vienna, Austria, 2022; pp. 32–58.
36. McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; Sonderegger, M. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 498–502.
37. McCloy, D.R. phonR: Tools for Phoneticians and Phonologists, R package version 1.0-7. 2016. Available online: https://www.r-project.org (accessed on 21 February 2024).
38. Verbrugge, R.R.; Strange, W.; Shankweiler, D.P.; Edman, T.R. What information enables a listener to map a talker's vowel space? *J. Acoust. Soc. Am.* **1976**, *60*, 198–212. [CrossRef] [PubMed]
39. Zekveld, A.A.; Kramer, S.E.; Festen, J.M. Cognitive Load During Speech Perception in Noise: The Influence of Age, Hearing Loss, and Cognition on the Pupil Response. *Ear Hear.* **2011**, *32*, 498–510. [CrossRef] [PubMed]

40. Baese-Berk, M.M.; Morrill, T.H. Speaking rate consistency in native and non-native speakers of English. *J. Acoust. Soc. Am.* **2015**, *138*, EL223–EL228. [CrossRef] [PubMed]
41. Ronnlund, M.; Nyberg, L.; Backman, L.; Nilsson, L.G. Stability, growth, and decline in adult life span development of declarative memory: Cross-sectional and longitudinal data from a population-based study. *Psychol. Aging* **2005**, *20*, 3–18. [CrossRef] [PubMed]
42. Summers, V.; Leek, M.R. F0 Processing and the Separation of Competing Speech Signals by Listeners With Normal Hearing and With Hearing Loss. *J. Speech Lang. Hear. Res.* **1998**, *41*, 1294–1306. [CrossRef] [PubMed]
43. Madsen, S.M.K.; Dau, T.; Oxenham, A.J. No interaction between fundamental-frequency differences and spectral region when perceiving speech in a speech background. *PLoS ONE* **2021**, *16*, e0249654. [CrossRef] [PubMed]
44. Stickney, G.S.; Assmann, P.F.; Chang, J.; Zeng, F.G. Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences). *J. Acoust. Soc. Am.* **2007**, *122*, 1069–1078. [CrossRef] [PubMed]
45. Garnier, M.; Dohen, M.; Loevenbruck, H.; Welby, P.; Bailly, L. The Lombard Effect: A physiological reflex or a controlled intelligibility enhancement? In Proceedings of the 7th International Seminar on Speech Production, Ubatuba, Brazil, 13–15 December 2006; pp. 255–262.
46. Cooke, M.; Lu, Y. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers). *J. Acoust. Soc. Am.* **2010**, *128*, 2059–2069. [CrossRef]
47. Pettinato, M.; Tuomainen, O.; Granlund, S.; Hazan, V. Vowel space area in later childhood and adolescence: Effects of age, sex and ease of communication. *J. Phon.* **2016**, *54*, 1–14. [CrossRef]
48. Cowley, C.M. The Effects of Distracting Background Audio on Speech Production. Master's Thesis, Brigham Young Universit, Provo, Utah, 2020.
49. Le, G.; Tang, Y. The Lombard Effect on the Vowel Space of Northern Vietnamese. In Proceedings of the 20th ICPhS, Prague, Czech Republi, 7–11 August 2023; pp. 1167–1171.
50. Derwing, T.M.; Munro, M.J. Accent, Intelligibility, and Comprehensibility: Evidence from Four L1s. *Stud. Second Lang Acquis* **1997**, *19*, 1–16. [CrossRef]
51. Miller, S.E.; Schlauch, R.S.; Watson, P.J. The effects of fundamental frequency contour manipulations on speech intelligibility in background noisea. *J. Acoust. Soc. Am.* **2010**, *128*, 435–443. [CrossRef]
52. Brenier, J.; Cer, D.; Jurafsky, D. The detection of emphatic words using acoustic and lexical features. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 3297–3300.