ON MEASURING THE INTELLIGIBILITY OF SYNTHETIC SPEECH IN NOISE — DO WE NEED A REALISTIC NOISE ENVIRONMENT?

Tuomo Raitio[1]*, Marko Takanen*[1]*, Olli Santala*[1]*, Antti Suni*[2]*, Martti Vainio*[2]*, and Paavo Alku*[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
[2]University of Helsinki, Institute of Behavioural Sciences, Helsinki, Finland

ABSTRACT

Assessing the intelligibility of synthetic speech is important in creating synthetic voices to be used in real life applications, especially for the ones involving interfering noise. This raises the question how to measure the intelligibility of synthetic speech to correctly simulate such conditions. Conventionally, this has been done using a simple listening test setup where diotic speech and noise are played to both ears with headphones. This is indeed very different from the real noise environment where speech and noise are spatially distributed. This paper addresses the question whether a realistic noise environment should be used to test the intelligibility of synthetic speech. Three different test conditions, one with multichannel reproduction of noise and speech, and two headphone setups are evaluated. Tests are performed with natural and synthetic speech, including speech especially intended for noisy conditions. The results indicate a general trend in all setups but also some interesting differences.

Index Terms— synthetic speech, speech in noise, intelligibility, multichannel reproduction, Lombard speech

1. INTRODUCTION

The intelligibility and quality of many text-to-speech (TTS) systems today are close to those of natural speech. However, in real applications speech synthesizers should be able to cope with adverse conditions with multiple noise sources and deliver the message through the interfering noise to the listener. This requires building special synthetic voices and assessing their intelligibility in different noise environments.

Conventionally, the intelligibility of synthetic speech in noise has been evaluated in simplified auditory environments. Usually, mixed speech and noise signal is delivered to the listener through headphones [1] or telephone [2]. Speech intelligibility in the presence of noise is obtained for that particular situation, but how well do those evaluation setups correspond to the noise environments confronted in real applications? Typically, spatially distributed sources of various types of noises are masking the synthetic voice, which is indeed different from the conventional evaluation setups.

This study addresses the question whether a more realistic noise environment should be used in assessing the intelligibility of synthetic speech. Although speech intelligibility in noise has been widely studied, as is shown in the next section, we are not aware of any studies specifically addressing this question. Moreover, in this study the perception of speech quality and suitability to the noise environment is assessed in various evaluation setups.

This research is supported by the Academy of Finland (projects 135003 LASTU programme, 1128204, 1218259, 121252), MIDE UI-ART, and Nokia Foundation. The authors would also like to thank E. Jokinen, M. Hiipakka, M.-V. Laitinen, and H. Pulakka of Aalto University for their help.

2. SPEECH INTELLIGIBILITY

Speech intelligibility depends on three main factors: (1) level and type of speech, (2) level and type of noise, and (3) acoustic environment. Naturally, the level and type of speech define the baseline for speech intelligibility. For example Lombard speech [3], i.e., speech that is produced in the presence of noise, can be more intelligible compared to normal speech even after loudness normalization [4]. Synthetic speech is usually less intelligible than natural speech due to artifacts in the speech signal and prosody. However, our recent studies indicate that it is also possible to create a synthetic (Lombard) voice that is more intelligible than natural speech [5, 6].

Secondly, speech intelligibility is affected by the level and type of noise. The relative levels of speech and noise, i.e., the signal-to-noise ratio (SNR), has naturally a major effect on intelligibility. The noise itself can be characterized by its spectrum, temporal structure, and the spatial distribution of its components. Both the spectrum and the temporal structure of noise cause auditory masking which decreases speech intelligibility [7]. The effect of the spatial distribution of noise to speech intelligibility is based on binaural hearing and thus the localization of sound sources. It has been shown that spatial separation between speech and noise enhances intelligibility. This effect has been shown for both broadband noise and competing speech sources [7, 8, 9, 10].

Thirdly, speech intelligibility is affected by the impulse response of the acoustic environment in which speech is reproduced. Long impulse responses will create reverberant sound that will decrease intelligibility, whereas strong early reflections may increase the SNR and thus increase intelligibility. The effect of the channel (e.g., loudspeakers, headphones, telephone) also affects intelligibility by introducing linear and nonlinear modifications to the signal.

There are numerous methods for testing the intelligibility of speech, but it is hardly possible to take into account all the factors in a single assessment. Speech intelligibility can be partly predicted based on the studies of each individual factor. However, speech intelligibility depends in a complex manner on the properties of the interfering signals, the number and spatial configuration of them, and the acoustic environment [8]. Therefore, an assessment between different evaluation scenarios is justified, especially for synthetic speech that may show different behavior compared to natural speech.

3. EXPERIMENTS

The effect of the sound reproduction setup on the speech intelligibility and quality was evaluated by conducting the same listening test using three different setups. More precisely, either a multichannel loudspeaker setup or one of two different headphone reproduction setups were used. The listening test was conducted in a listening

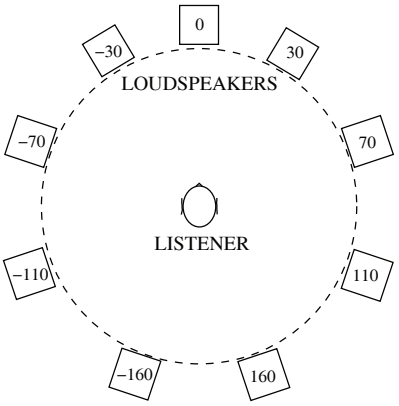

Fig. 1. Illustration of the loudspeaker setup in the multichannel test. Nine identical loudspeakers were positioned around the listener (angles depicted for each loudspeaker) in order to create a realistic noise environment.

room fulfilling the requirements of ITU-R BS.1116-1. The average reverberation time of the room is 0.3 s.

In the listening test, the listeners were presented with speech samples in the presence of masking noise using two different SNR levels. The speech samples consisted of natural and synthetic versions of normal and Lombard speech. Additionally, two different noise types were used. The task of the listener was to type in the heard sentence and rate the quality as well as the suitability of the speech samples. Speech intelligibility was then evaluated based on the word error rate (WER).

3.1. Multichannel Evaluation Setup

The multichannel evaluation setup consisted of nine identical loudspeakers (Genelec 8260A) positioned at 2.4–2.6 meter distances around the listener. The loudspeaker responses at the listener position were equalized using DSP. The loudspeaker setup is illustrated in Fig. 1. The loudspeaker directly at the front was used to reproduce the speech samples and all of the nine loudspeakers were employed in the reproduction of the masking noise. B-format recordings of the noise were used, consisting of four channels: W, X, Y and Z. The signals to the nine loudspeakers of the noise stimuli were obtained by rendering the B-format microphone recording with Directional Audio Coding [11].

3.2. Headphone Evaluation Setups

Two different scenarios of headphone reproduction using circumaural headphones (Sennheiser HD580) were employed, namely, diotic (identical noise to both ears) and dichotic (different noise to both ears). These setups are referred to as mono and stereo setups, respectively. In both of these scenarios, speech samples were reproduced diotically. The diotic noise scenario (mono setup) was generated by feeding the signal from the omnidirectional (W) channel of the B-format microphone recording to both channels of the headphones.

The dichotic noise scenario (stereo setup) was generated by feeding a stereo noise signal to the left and right channels of the headphones. The noise signal was created from the W and Y channels of the B-format microphone, corresponding to a stereophonic recording with two cardioid microphones facing the directions of $\pm 90^\circ$.

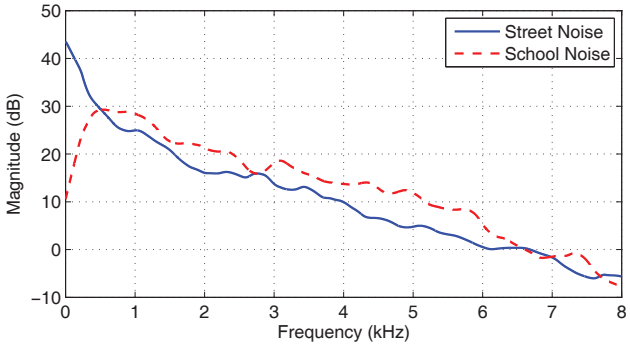

Fig. 2. Spectra of street and school noises used in the test.

The sound reproduction levels between the loudspeaker and the headphone setups were equalized by employing in-ear microphone measurements with a human subject, and normalizing the A-weighted sound pressure levels (SPLs) of the signals in front of the ear canal.

3.3. Noise Material

Two types of noise were used in the listening test: street and school noise. Street noise has most of the energy at low frequencies while school noise has relatively more energy at the higher frequencies. The spectra of the two noises are shown in Fig. 2. Both noises were reproduced with two A-weighted SPLs: moderate (63 dB) and loud (70 dB). The average SNRs of speech and noise were -1 dB and -8 dB, respectively.

3.4. Speech Material

Four types of speech signals were used in the listening tests by involving both natural and synthetic sentences that correspond to two speaking styles (normal and Lombard speech). All speech material originates from a Finnish male speaker, whose normal and Lombard speech was recorded [6] and further used to train a hidden Markov model (HMM) based speech synthesizer as described below. The notations used for the four test speech types are given in Table 1.

Statistical parametric TTS system GlottHMM [12] was used to build the synthetic voices. GlottHMM was chosen since it has been shown to be able to reproduce Lombard characteristics in synthetic speech [5, 6]. Previously, we have compared different methods for synthesizing Lombard speech [6], and the best method according to the comparison was selected for this study. In this method, recorded Lombard speech is used for adapting normal speech models, and then extrapolation is used between the normal and adapted speech models in order to get stronger Lombard characteristics to the synthetic voice.

The active speech level, or loudness, of all speech samples was normalized using the method in ITU-T P.56. The A-weighted SPLs of the normalized speech samples are shown in Table 1.

3.5. Listening Tests

In the listening test, test subjects were presented with speech samples masked by noise, both either from loudspeakers or headphones depending on the evaluation setup (presented in random order). A representative set of 144 short Finnish sentences (average length of 12.7 syllables) designed for intelligibility tests [13] was presented in

random order to each listener. The task of the listener was to type in what he or she heard as accurately as possible, and WERs were evaluated over the sentences in each condition, taking separately into account the inflectional and derivational suffixes. After this task, the test subject was allowed to listen to the sample as many times as he or she liked, and rate the speech sample according to two questions: *How would you rate the quality of the speech sample?* and *How suitable was the speaking style considering the noise environment?* A continuous scale from 0 to 100 with the following verbal descriptions was used for both questions: *bad (0) – poor (25) – fair (50) – good (75) – excellent (100).*

3.6. Results

Seventeen native speakers of Finnish (15 male and 2 female) between 18 and 35 years of age with no known hearing impairments participated in the listening test. The results of the test were analyzed using a five-way analysis of variance (ANOVA) with the evaluation setup ($setup$), type of speech ($speech$), type of noise ($noise$), and SNR level (snr) as fixed variables and the listener as a random variable. The ANOVA analysis was performed separately for intelligibility, quality, and suitability ratings using the same 5% significance level. The marginal means and their 95% confidence intervals were computed and Dunnett's T3 post hoc test with the significance level of 5% was applied to gain more insight about the nature of the effects. Due to the lack of space only the effects that were found significant and are interesting considering the scope of this paper are analyzed with more detail.

3.6.1. Intelligibility

The word error rates with 95% confidence intervals for each evaluation setup are shown in Fig. 3. All the effects of the fixed variables and their interactions were found significant by ANOVA except for $setup \times speech$, $setup \times noise \times speech$, $setup \times noise \times snr$, and $setup \times speech \times snr$. The overall WER was considerably better in the case of stereo setup compared to mono and multichannel setups that share the same overall WER [$setup$: F(2,32) = 41.98, $p \ll 0.001$]. Additionally, different reproduction techniques had similar WER in high SNR situation, but not in low SNR situation [$setup \times snr$: F(2,32) = 16.79, $p < 0.001$]. Also natural and synthetic samples of Lombard speech differed in terms of WER in the case of loudspeaker reproduction of school noise with low SNR [$setup \times speech \times noise \times snr$: F(6,96) = 2.60, $p < 0.05$]. The aforementioned findings were confirmed with the post hoc test.

3.6.2. Quality

The overall results for quality rating with 95% confidence intervals are shown in Fig. 4. All the main effects of the fixed variables and their two-way interactions with snr were found significant by ANOVA. The quality rating for mono setup was considerably lower than the ones for stereo and multichannel setups [$setup$: F(2,32) = 7.47, $p < 0.01$]. The differences between the reproduction methods were significant only in the case of low SNR [$setup \times snr$: F(2,32) = 11.57, $p < 0.001$]. The post hoc test confirmed these findings.

Table 1. *Test voice types and their averaged A-weighted SPLs after loudness normalization with ITU-T P.56.*

Type	Description	SPL
nat_norm	Natural normal speech	59 dB
syn_norm	Synthetic normal speech	61 dB
nat_lomb	Natural Lombard speech	63 dB
syn_lomb	Synthetic Lombard speech	63 dB

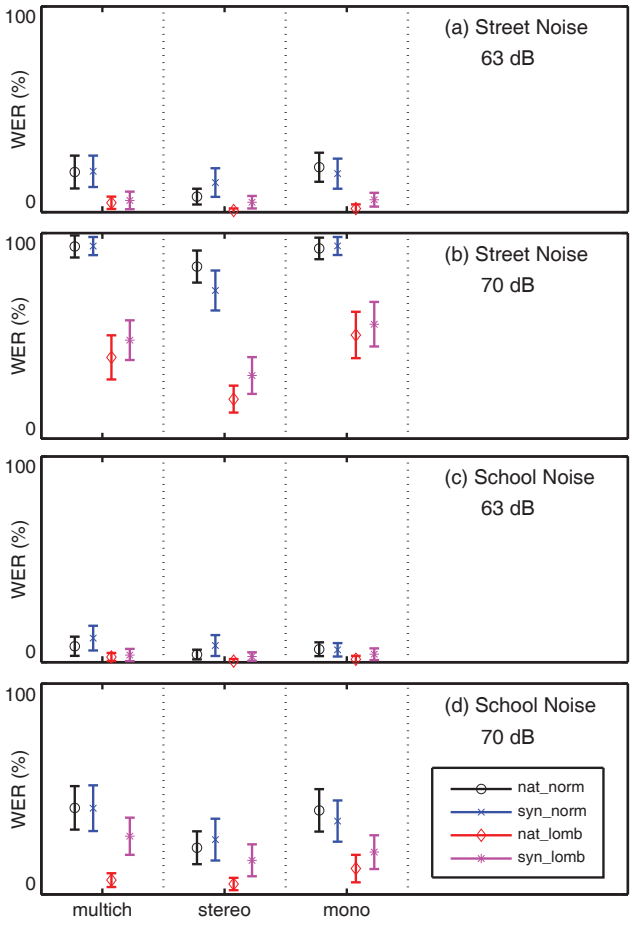

Fig. 3. Word error rates for different types of speech in street noise at (a) 63 dB and (b) 70 dB, and in school noise at (c) 63 dB and (d) 70 dB for each evaluation setup: multichannel, stereo, and mono.

3.6.3. Suitability

The overall results for suitability rating with 95% confidence intervals are shown in Fig. 4. ANOVA returned significant effects for all the fixed variables and their interactions except for $setup \times speech$, $setup \times snr$, and $setup \times noise \times snr$. The suitability ratings differed between reproduction setups the average rating being highest for stereo and lowest for mono setup [$setup$: F(2,32) = 12.37, $p < 0.001$]. Moreover, natural and synthetic samples of Lombard speech samples differed from each other in terms of suitability only in the cases of loudspeaker reproduction of school noise with low SNR, and in high SNR situations with stereo setup [$setup \times speech \times noise \times snr$: F(6,96) = 3.03, $p < 0.01$]. These findings were also confirmed with the post hoc test.

4. DISCUSSION AND CONCLUSIONS

The results of the experiment show a general trend in all of the evaluation setups, according to which both natural and synthetic Lom-

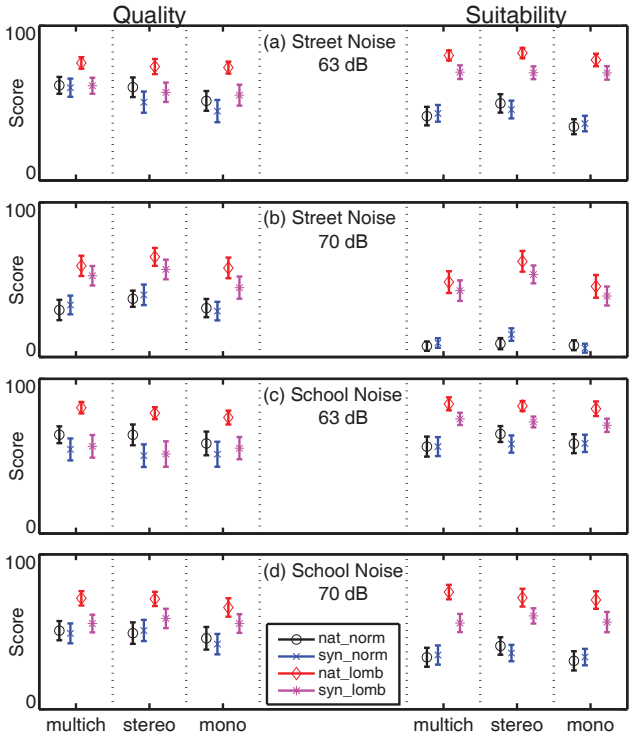

Fig. 4. Results of the subjective evaluation: quality and suitability of different types of speech in street noise at (a) 63 dB and (b) 70 dB, and in school noise at (c) 63 dB and (d) 70 dB for each evaluation setup: multichannel, stereo, and mono.

bard speech are statistically more intelligible than normal speech. In addition, word error rates were on average higher in loud noise compared to moderate noise, and in street noise compared to school noise. The latter is explained by the more effective masking of the low-frequencies of the male voice by the street noise.

However, interesting differences between the evaluation setups were also found. First, the stereo setup gave on average 9 percentage points lower WERs compared to both multichannel and mono setups. The intelligibility difference between mono and stereo setups is in accordance with the literature, but the similarity of the mono and multichannel setups was unexpected. The higher WERs of the multichannel setup may be due to the response of the listening room that may lower intelligibility, although the reverberation time of the room was relatively low. Alternatively, the lower WER scores of the stereo setup may stem from the nature of the constructed noise; since it is created by using only the W and Y channels of the B-format microphone, some of the noise present in the other two cases was absent. This causes some differences in the presented noise that may increase intelligibility despite the equal SPLs.

Second, the mono setup gave on average lower quality scores than the other two setups. With headphone listening, the artifacts of synthetic speech are perceived easier compared to loudspeaker reproduction, but the lower quality ratings were not confined only to synthetic speech samples. Although the subjective quality rating was intended to describe only the quality of the speech samples, disregarding the noise, people tend to score speech samples higher in quality in high SNR than in low SNR. Thus, the lower quality ratings in the mono setup may indicate some other phenomenon.

Third, the speech sounds were considered least suitable in the mono setup, and most suitable in the stereo setup. In addition, the evaluation setup also affected the ranking of different speech types in the presence of school noise with low SNR; natural and synthetic samples of Lombard speech differ in terms of WER in the case of loudspeaker reproduction, but not in other setups.

Although none of the evaluation setups could be considered better or worse compared to each other, the study shows that there are differences between them. Finally, the purpose and type of the study may define what type of evaluation is to be used; for mass listening tests headphone listening yields good results, but for more specific testing, more realistic test setups may be beneficial. Thus, the answer to the question of the title might be that a realistic noise environment is not a requisite for intelligibility testing of synthetic speech in noise, but it may yield additional information that other setups might miss.

5. REFERENCES

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2010," in *The Blizzard Challenge 2010 workshop*, http://festvox.org/blizzard.

[2] B. Langner and A. W. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proc. ICASSP*, 2005, pp. 265–268.

[3] W. Van Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[4] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[5] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *The Blizzard Challenge 2010 workshop*, 2010, http://festvox.org/blizzard.

[6] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-based Lombard speech synthesis," in *Proc. Interspeech*, 2011, pp. 2781–2784.

[7] J. Blauert, *Spatial Hearing*, The MIT Press, Cambridge, MA, USA, revised edition, 1997.

[8] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acust. unit. Acust.*, vol. 86, no. 1, pp. 117–128, 2000.

[9] R. Drullman and A. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *J. Acoust. Soc. Am.*, vol. 107, no. 4, pp. 2224–2235, 2000.

[10] M. Hawley, R. Litovsky, and J. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and article of interferer," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843, 2004.

[11] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[12] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[13] M. Vainio, A. Suni, H. Järveläinen, J. Järvikivi, and V.-V. Mattila, "Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish," *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1742–1750, 2005.