Using linguistic predictability and the Lombard effect to increase the intelligibility of synthetic speech in noise

Cassia Valentini-Botinhao, Mirjam Wester

The Centre for Speech Technology Research, University of Edinburgh, UK

cvbotinh@inf.ed.ac.uk, mwester@inf.ed.ac.uk

Abstract

In order to predict which words in a sentence are harder to understand in noise it is necessary to consider not only audibility but also semantic or linguistic information. This paper focuses on using linguistic predictability to inform an intelligibility enhancement method that uses Lombard-adapted synthetic speech to modify low predictable words in Speech Perception in Noise (SPIN) test sentences. Word intelligibility in the presence of speech-shaped noise was measured using plain, Lombard and a combination of the two synthetic voices. The findings show that the Lombard voice increases intelligibility in noise but the intelligibility gap between words in a high and low predictable context still remains. Using a Lombard voice when a word is unpredictable is a good strategy, but if a word is predictable from its context the Lombard benefit only occurs when other words in the sentence are also modified.

Index Terms: intelligibility enhancement, speech in noise, HMM-based speech synthesis, SPIN test

1. Introduction

Intelligibility of Text-to-Speech (TTS) systems degrade considerably in noise. Although research aimed at generating highly intelligible synthetic speech has been carried out [1, 2, 3, 4, 5], using linguistic information to selectively enhance words has not been taken into account yet. Current modifications rely mainly on the audibility of speech in noise, often guided by objective measures of intelligibility [6, 7, 8] based on the acoustic and effective processing of the human auditory system. Usually, modifications are applied to the whole sentence, but possibly they should not be "on" the whole time. It would be better if the degree of modification was guided by a trade-off between intelligibility and naturalness/quality of the speech.

Valentini-Botinhao et al. [9] explored the idea of using word-level intelligibility predictions to selectively boost the harder-to-understand words in a sentence, aiming to improve overall intelligibility in the presence of noise. First, the intelligibility of a set of words from dense and sparse phonetic neighbourhoods was evaluated in isolation. The resulting intelligibility scores were used to inform two sentence-level experiments. In the first experiment, the signal-to-noise ratio of one word was boosted (in terms of energy reallocation) to the detriment of another word. In general, the sentence-level intelligibility did not improve. The intelligibility of words in isolation and in a sentence were found to be significantly different, both in clean and in noisy conditions. For the second experiment in [9], one word was selectively boosted while slightly attenuating all other words in the sentence. This strategy was successful for words that were poorly recognised in that particular context. However, a reliable predictor of word-in-context intelligibility remains elusive, since this involves – as the results in [9] indicated – semantic, syntactic and acoustic information about the word and the sentence.

In this paper, rather than exploiting the acoustic confusability of words we create strategies for intelligibility enhancement that consider linguistic predictability instead. As early as 1951, Miller et al. [10] showed that sentence context imposes constraints on the set of alternative words that are available as responses at particular locations in the sentence. The finding that the predictability of words in a sentence influences their intelligibility has repeatedly been shown in a variety of studies (e.g., [11, 12, 13]).

In the current experiments, we vary the degree of linguistic predictability by using the Speech Perception in Noise (SPIN) test. The SPIN test was originally developed as an assessment tool to examine how (hearing-impaired) listeners utilize linguistic and acoustic information in a sentence [14]. Because the SPIN test was designed to be a test of everyday speech perception the target words were placed in a sentence and speech babble was used as noise. The test contains two types of sentences, high-predictable (HP) and low-predictable (LP). In an HP sentence, the final word is predictable from the semantic content of the sentence whereas in an LP sentence it is unpredictable. Thus, it is a more realistic task than, for instance, semantically unpredictable sentences (SUS) which are often used in speech synthesis evaluation [15] and in which all the words are chosen to be semantically unpredictable. One of the disadvantages of using the SPIN set rather than the SUS set is that each sentence only gives one datapoint as only the final word is usually transcribed. This can be partially ameliorated by asking listeners to transcribe the whole sentence and use the datapoints for the other words as well. It has been shown that the recognition rates of the final word are not affected by altering the task from final word transcription to full sentence transcription [16, 17].

A previous study looking at using SPIN for speech synthesis evaluation [16] showed that the intelligibility of LP words is quite low (25–35 % WER). These are exactly the type of words that Valentini-Botinhao et al. [9] showed benefit from boosting. Based on the findings in [9], we expect that applying a modification to the final LP words will lead to larger improvements in WER than boosting the final HP words.

The type of modification we are exploring in this study is Lombard synthetic speech. Lombard speech is mainly characterised by increases in F_0, intensity, and duration [18, 19, 20, 21, 22]. We compare the intelligibility of synthesised SPIN sentences in speech-shaped noise using plain speech, Lombard speech and a combination of both within a sentence. In the combination condition, the final word of the sentence is in a Lombard voice while the rest is produced using plain speech.

The key questions we are interested in answering in this pa-

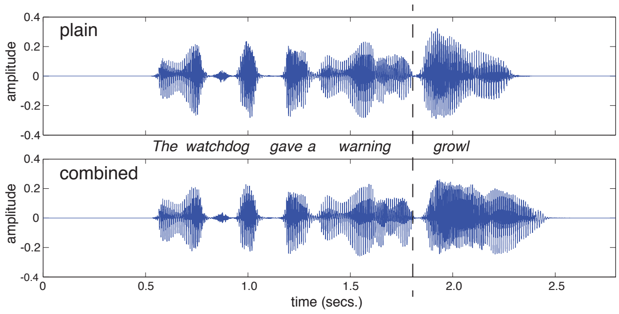

Figure 1: Waveform of a sentence generated by the plain (top) and the combined (bottom) strategy.

per are: Can we improve the intelligibility of a sentence (or just the final word) by selectively modifying the final word? What is the effect of switching between plain and Lombard speech on the intelligibility of the other words in the sentence? The remainder of this paper sets out the details of how our stimuli were created, followed by a description of the listening experiment. Next, the results are presented and finally a discussion of the findings is given.

2. Method

In our experiments, we use two synthetic voices, one trained on plain speech data (plain voice) and the other trained on Lombard –speech produced in noise– data (Lombard voice). In order to examine the efficacy of a strategy in which there is a switch from plain to Lombard for the last word in a sentence, we compare the following modification strategies:

- plain: use the plain voice to generate the whole sentence;
- Lombard: use the Lombard voice to generate the whole sentence;
- combined: use the plain voice to generate the first part of the sentence and Lombard to generate the last word.

2.1. Speech material

The plain voice was created from a high quality average voice model which was adapted using a three hour recording of a British male speaker as described in [5]. We used a hidden semi-Markov model as the acoustic model.

To create the samples in the combined style, we construct a model sequence by concatenating model parameters from the plain and the Lombard voice. First, we extract the model sequence corresponding to the whole sentence for both the plain and the Lombard voice. We then combine these two sequences: the first part, i.e., the model sequence corresponding to the whole sentence up until the last word, is copied from the plain model sequence and the second part, the last word, is taken from the Lombard model sequence. We then use the maximum likelihood parameter generation algorithm [23] to find a smooth trajectory of acoustic parameters from this combined model sequence. In this way, any artefact that could arise from switching from one style to another is smoothed out during parameter generation. Fig. 1 shows an example of a sentence generated by the plain and the combined modification strategy. Acoustic samples of the three conditions are available at: `wiki.inf.ed.ac.uk/CSTR/SpinLombard`

2.2. Noise material

To ensure the task would be challenging enough, we added speech to stationary noise – speech-shaped noise built from recordings of a female speaker – at -3 dB signal-to-noise ratio (SNR). This SNR value was chosen such that word accuracy was 50 % on average [16]. The SNR of the last word was kept fixed across modifications as well as the SNR of the first section of the sentence, so that any benefit found for the Lombard voice would not be due to an increase in SNR.

2.3. SPIN sentences

The full SPIN test comprises 400 sentences, organized in eight forms of 50 sentences each [14]. Half of the sentences are characterized as high-predictable (HP) and the rest as low-predictable (LP). In HP sentences, the final word (key word) is predictable from the semantic context of the sentence (e.g., “The story had a clever plot”) whereas in an LP sentence the final word is unpredictable (e.g., “You’re discussing the plot”). In HP sentences, the context words that help predict the key words are called pointer words. Each list is paired with another list such that the key words in HP sentences in one list are in LP sentences in the other list. All sentences are constrained to contain five to eight words and all key words are monosyllabic. The sentences are balanced for intelligibility, key word familiarity and predictability and phonetic content [14]. A later study by Clarke [24] showed that the key word neighbourhood ratios are also equivalent. Percent correct transcription of the final word is the conventional measure of listener performance.

2.4. Listening experiment

The listening test comprises the first four forms of the SPIN test, a total of 200 sentences. Each participant listened to two of the forms, either 1 and 3 or 2 and 4. This was to ensure that each key word was heard only once by a participant. Listeners typed what they heard, after which the following stimulus was presented. They were instructed to type ’x’ if they could not make out the word(s). To facilitate the presentation, the sentences were organized in four blocks of 25. We balanced the modification type (plain/Lombard/combined) across listeners. This means that it took six listeners to hear all 200 sentences in each of the 3 styles. Prior to the experiment, a small hearing test was carried out for each participant to exclude data from listeners with poor hearing. The listening test started with a short practice session to familiarize listeners with the task. The results of this paper are based on data from 42 participants. All participants were native English speakers.

3. Results

To calculate word accuracy levels, we compared the transcriptions provided by participants with the correct scripts used to generate the test, considering not only misspellings but also different spellings of the same word and homophones. We present results organized in word groups: the key word (final word) results are presented in Fig. 2, the other words in the sentence in Fig. 3 and the pointer words in Fig. 4.

Word accuracy is presented as the percentage of correctly recognised words.. These percentages are calculated per listener and then averaged across listeners. Statistical analysis is represented here in notched boxplots. The median is displayed as a solid line and the notches represent confidence intervals: if the intervals of two medians do not overlap, they are signifi-

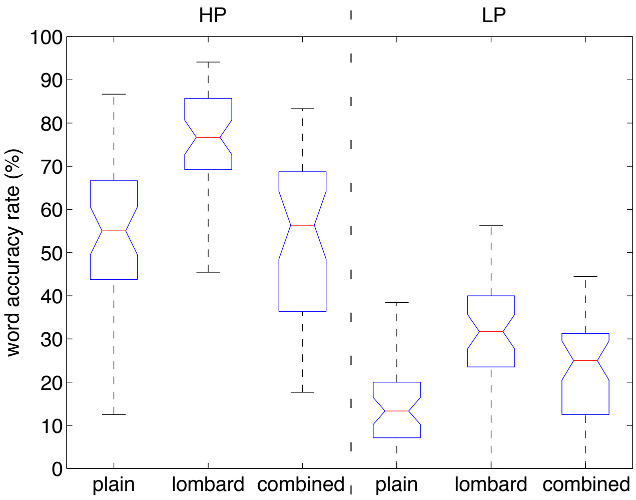

Figure 2: Word accuracy results for the key words in HP and LP context for plain, Lombard and combined styles.

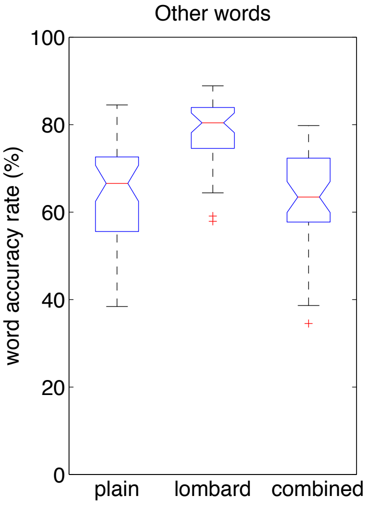

Figure 3: Word accuracy results for the other words in the sentence for plain, Lombard and combined styles.

cantly different at the 5 % significance level. The top of the box (upper-quartile) marks the 75th percentile for the data set and the bottom (lower-quartile) the 25th percentile. The whiskers show the highest and lowest values that are within 1.5 times the interquartile range (IQR) of the box edges. Those values are not considered outliers.

Fig. 2 presents the accuracy rates for key words in HP (high-predictable) and LP (low-predictable) contexts. The results are organized by modification type. Looking first at the results obtained with the plain voice we can see a significantly large intelligibility gap for a word produced in a low and in a highly predictable context: from 14.7 % to 53.9 %. The results also show that the Lombard voice is more intelligible than the plain voice both in HP and LP contexts. In the HP context, the plain voice reached an average intelligibility rate of 53.9 % while the Lombard voice obtained 76.3 %, a relative increase of 42 %. When the last word is harder to predict, the LP context, the Lombard advantage is even larger: from 14.7 % to 31.9 %, a relative increase of 116 %. The Lombard benefit is however not large enough to bridge the gap caused by the lack of linguistic predictability, as we still see a significant difference between scores in LP and HP contexts.

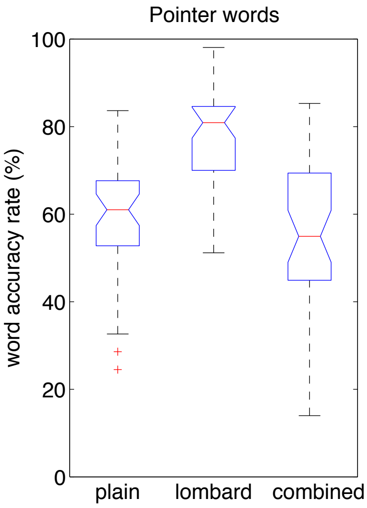

Figure 4: Word accuracy results for the pointer words in HP sentences for plain, Lombard and combined styles.

The strategy of combining the plain and Lombard styles improves intelligibility of the last word only when that word is in an unpredictable context. In the HP context, the combined condition results in an average accuracy of 53.6 % while in the LP context the combined strategy reaches 23.6 %, a relative increase of 60 % over the plain voice. In the HP context, the Lombard and the combined style results are not significantly different, although there is an advantage of the Lombard style over the combined style.

Fig. 3 shows accuracy scores for the other words in the sentence, that is, excluding the key word. As expected the plain and the combined voices result in similar intelligibility scores as both strategies use the same models to generate these words. The average word accuracy across listeners was 64.5 % for plain and 62.4 % for combined. The intelligibility rate of words produced by the Lombard voice was 78.3 %, a relative increase of 21 % over the plain voice.

For HP sentences, i.e., the last word is highly predictable, certain words in the sentence can be classified as pointer words as they increase the predictability of the key word. The ranking of the scores for pointer words in HP sentences, shown in Fig. 4, is similar to that seen for all other words: plain and combined are not significantly different (59.4 % and 56.0 %) and once again the Lombard voice resulted in a higher intelligibility rate (78.5 %), a relative increase of 32 %.

4. Discussion

Results indicate that the Lombard voice is more intelligible than the plain voice: intelligibility benefits are clear across all words in a sentence and for the final word, intelligibility gains are seen whether the word is highly predictable or not. The relative benefit over the plain voice is much higher when the final word is harder to predict, i.e., when the listener only has acoustic cues to rely on to understand the word. In this situation, the Lombard strategy is especially beneficial as it aims to enhance such acoustic cues. However, the Lombard modification alone is not enough to overcome the intelligibility loss that the lack of linguistic predictability creates, which illustrates how important it is to control the strength of a modification accordingly.

When the final word is generated using the Lombard model but the rest of the sentence is generated using the plain model the Lombard benefit ceases to apply if the last word can be guessed from its context. In this condition, the combined voice does not seem to improve the intelligibility of the last word even though the last word is spoken in a Lombard style, which we have seen can lead to intelligibility gains in that context if applied across the whole sentence. This indicates that enhancing the last word by using a Lombard voice does not improve intelligibility and that the benefit seen when using the Lombard voice actually resulted from the fact that the pointer words in the sentence were better understood. When there is no semantic context to help the understanding of the final word, applying an intelligibility enhancement strategy to that word alone can be beneficial, as the results of the key word in the LP context show.

Even though keeping the Lombard style across the whole sentence might yield higher intelligibility scores across all words, when a word is highly predictable maintaining the Lombard style might not be necessary. A strategy that keeps track of a word's linguistic predictability could be a good compromise between the loss in naturalness and intelligibility requirements. (A loss in naturalness and quality has been shown for synthetic Lombard speech [25].) Switching between speaking styles, which can be considered equivalent to switching an intelligibility enhancement strategy on and off, does not significantly decrease intelligibility scores of words in low predictability contexts. In other words, a strategy that fluctuates between unmodified and modified speech depending on a word's linguistic predictability can potentially maintain the same intelligibility level as a strategy where modification is on permanently.

5. Conclusions

In this paper, we evaluated whether a speech intelligibility enhancement strategy can make use of word predictability. To this end, we used the set of SPIN sentences which was specifically designed to test intelligibility of a word, in this case the last word of the sentence, in a low and in a high predictability context. To assess the effect that a synthetic Lombard voice has in these two different contexts we evaluated three modification strategies: plain, Lombard and combined, where combined involves changing from plain to Lombard only at the last word. Results show that even though the Lombard voice increases intelligibility in noise, the gap between a low and high probable context was not closed, illustrating the need for even stronger modifications when linguistic predictability is low. Switching between plain and Lombard is as good as keeping Lombard on the whole time when the last word is not easy to predict from the context. When the key word is highly predictable switching between plain and Lombard is not as effective, as the pointer words need to be enhanced to give the full benefit of Lombard to the key word. Future work includes using predictions from a language model to guide the interpolation factor between plain and Lombard models and we will also investigate controlling other sorts of intelligibility enhancement modifications.

6. Acknowledgements

This research was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

7. References

[1] B. Langner and A. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proc. ICASSP*, vol. 1, 18-23, 2005, pp. 265–268.

[2] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-based Lombard speech synthesis," in *Proc. Interspeech*, Florence, Italy, August 2011.

[3] B. Picart, T. Drugman, and T. Dutoit, "Continuous control of the degree of articulation in HMM based speech synthesis," in *Proc. Interspeech*, Florence, Italy, 2011.

[4] M. Nicolao, J. Latorre, and R. Moore, "C2H A computational model of H&H-based phonetic contrast in synthetic speech," in *Proc. Interspeech*, Portland, USA, September 2012.

[5] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise," in *Proc. Interspeech*, Portland, USA, September 2012.

[6] M. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[7] C. Christiansen, M. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7-8, pp. 678–692, 2010.

[8] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, Dallas, USA, March 2010, pp. 4214–4217.

[9] C. Valentini-Botinhao, M. Wester, J. Yamagishi, and S. King, "Using neighbourhood density and selective SNR boosting to increase the intelligibility of synthetic speech in noise," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 133–138.

[10] G. Miller, G. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials." *Journal of Experimental Psychology*, vol. 41, no. 5, p. 329, 1951.

[11] K. Hutchinson, "Influence of sentence context on speech perception in young and older adults," *Journal of Gerontology*, vol. 44, no. 2, pp. P36–P44, 1989.

[12] M. Fallon, S. Trehub, and B. Schneider, "Children's use of semantic cues in degraded listening environments," *The Journal of the Acoustical Society of America*, vol. 111, p. 2242, 2002.

[13] A. Bradlow and J. Alexander, "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners," *The Journal of the Acoustical Society of America*, vol. 121, p. 2339, 2007.

[14] D. Kalikow, K. Stevens, and 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, p. 1337, 1977.

[15] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard challenge 2008," in *Blizzard Challenge Workshop*, 2008.

[16] P. Drakopoulou-Kalantzi, "An investigation on the intelligibility of synthetic speech by reviewing the SPIN test," Master's thesis, School of Informatics, University of Edinburgh, 2013.

[17] S. Gordon-Salant and P. Fitzgibbons, "Selected cognitive factors and speech recognition performance among young and elderly listeners," *Journal of Speech, Language and Hearing Research*, vol. 40, no. 2, p. 423, 1997.

[18] E. Lombard, "Le signe d'élévation de la voix [the sign of the elevation of the voice]," *Annales des maladies de l'oreille et du larynx*, vol. 37, pp. 101–119, 1911.

[19] W. Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analysis," *Journal of the Acoustical Society of America*, vol. 84, pp. 917–928, 1988.

[20] J. C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[21] R. Patel and K. Schell, “The influence of linguistic content on the Lombard effect,” *Journal of Speech, Language and Hearing Research*, vol. 51, p. 209, 2008.

[22] Y. Zhao and D. Jurafsky, “The effect of lexical frequency and Lombard reflex on tone hyperarticulation,” *Journal of Phonetics*, vol. 37, no. 2, pp. 231–247, 2009.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis,” in *Proc. ICASSP*, 2000, pp. 1315–1318.

[24] C. Clarke, “Lexical neighborhood properties of the original and revised speech perception in noise (SPIN) tests,” Indiana University, Research on Spoken Language Processing Progress Report 24, 2000.

[25] S. Palmaz López-Peláez, “Speech synthesis reactive to dynamic noise environmental conditions,” Master’s thesis, PPLS, University of Edinburgh, 2013.