Speech synthesis reactive to dynamic noise environmental conditions

Susana Palmaz López-Peláez[1], *Robert A. J. Clark*

Center for Speech Technology Research, The University of Edinburgh

susana.palmaz@nuance.com, rob.clark@ed.ac.uk

Abstract

This paper addresses the issue of generating synthetic speech in changing noise conditions. We will investigate the potential improvements that can be introduced by using a speech synthesiser that is able to modulate between a normal speech style and a speech style produced in a noisy environment according to a changing level of noise. We demonstrate that an adaptive system where the speech style is changed to suit the noise conditions maintains intelligibility and improves naturalness compared to traditional systems.

Index Terms: parametric speech synthesis, speech in noise, speech modulation, reactivity, listening evaluation

1. Introduction

It is well known that noise is detrimental to speech perception, especially to intelligibility. Humans compensate for the effect of noise by modulating their speech style to enhance perceptibility. In the case of synthesised speech, several attempts have been made to model this behaviour [1, 3, 4, 12]. Particularly in the case of statistical parametric speech synthesis, the possibility to manipulate specific speech features with higher flexibility has offered a more robust approach to generate different speech styles. Also, latest developments using these methods have produced an implementation of a synthesis engine that can manipulate speech parameters in real time [20]. The combination of these two factors has highlighted the need for an evaluation of the real benefits that can be obtained by using real-time modulated synthetic speech according to the listening conditions of the environment. This paper presents a listening evaluation with participants that compares the performance of an *oracle* speech synthesis system that simulates reactivity to noise with two non-reactive systems. The investigation presented in this report is a selection of the results achieved as a result of the first author's MSc dissertation at the University of Edinburgh.

The speech style that humans produce under the influence of noise is called the Lombard effect or reflex. It can be defined as the involuntary modulation that speakers do to improve their audibility in noise. Apart from an increase in intensity, the main acoustic characteristics that have been observed [1, 5, 6, 7] for this style are: increased pitch and pitch range, increased vowel duration, decreased speech rate, energy shift from low frequency bands to middle and high bands and shift in formant centre frequencies.

These changes are introduced as a strategy to reallocate the speech signal energy concentrations to frequency regions where there is less concentration of noise to reduce the effect of the masking. These studies have also shown how the modulation can present both inter- and intra-speaker variability. Furthermore, the Lombard reflex modulation is not binary in nature, but it is rather produced along a continuum of complex interactions between the speakers' speech production characteristics, the linguistic context and the environmental conditions [7].

In a real world scenario the noise conditions in which a TTS systems are used are rarely stable. Picking a neutral system or a system specifically trained for a noisy environment might not be sufficient to achieve a performance level that is consistently natural and intelligible. We look at noise not as a binary feature (present or absent across a whole utterance), like most previous studies, but as a dynamically changing constraint that affects the perceived quality of speech synthesis.

In this report we will focus on the introduction of speech production modifications to achieve maximum communicability, although speakers have at their disposal other resources to make their output more understandable, for example, simplifying the content of the message, adding linguistic redundancy, and using non-verbal back-channels [8].

Noise is extremely detrimental to the intelligibility of speech in general and in particular to speech synthesis. Just increasing the intensity to solve this problem is not a sufficient solution due to the creation of distortions [9] that can be magnified by the increase in the general levels of the the listening conditions. As the literature shows, humans do not rely exclusively on intensity modifications to overcome the impact of noise. In the same way, several methods have used more complex approaches to reproduce the acoustic features of Lombard speech [3, 9, 12], with different levels of success. Despite all efforts, it is not yet agreed how these different acoustic parameters contribute to intelligibility. There have also been more thorough research projects [8] and large scale evaluations and challenges [10, 11] that target the problem of intelligibility in noise.

Other avenues of research have tried to model Lombard speech directly to improve intelligibility. Concatenative methods [1, 12] lack the necessary flexibility to be moulded to specific noise environments further than using recordings of speech performed in the target noise. The main issue is that the elicitation of speech in noise is difficult to achieve with the consistency and productivity needed to create a high quality unit selection database. Speech in noise is not shouted speech, but it is still produced under effort, and cannot be maintained for long. Signal manipulation is also possible, and some concatenative methods [1] tried to used GMM transformations using only a small set of Lombard recordings. A different technique was used in [12] where the unit cost selection algorithm was modified to include a measure of intelligibility based on the Speech Intelligibility Index (SII), calculated offline. If the unit database is big enough, this method is be able to find the most intelligible unit. These methods demonstrated improvements over traditional systems.

Statistical parametric approaches are more flexible and rely on model manipulation, mainly adaptation, from an average model using a small set of recordings of Lombard speech. Us-

[1] Currently at Nuance Communications

ing widely accessed model transformation methods like MLLR and MAP they have offered successful results [13, 15]. Other methods modify selected parameters individually to match the modulation of Lombard speech [3, 4, 9, 14]. These methods do not require additional recordings or retraining, and they are focused instead on modifying and controlling the synthesis parameters in the vocoder directly. An interesting approach is given in [4], where the spectral shape is controlled according to the specific type and level of noise using the Glimpse Proportion Measure as an intelligibility measure to optimise the modification values for the cepstral coefficients. Finally, approaches based on model interpolation have also been evaluated through transcription tasks and subjective mean opinion score evaluations with positive results [3].

All the approaches presented above have been successful in offering intelligibility and naturalness improvements under noise, but these do not necessarily hold when the same system is used in quiet. We argue that it could be beneficial to be able to access both neutral and Lombard models depending on the presence/absence and acoustic characteristics of noise. This can only be achieved in a reactive framework that can control the synthesised output in real time.

Attempts like those outlined under the "PRESENCE" theory of speech [18, 19] propose to control intelligibility in real time using a speech recognition module as a means of automated feedback system. It is yet to be seen what real improvements can be achieved with this theoretical framework with an evaluation of a fully developed system under this method. The second approach to TTS reactivity is that offered in [20] in which a new framework was developed for HTS, namely pHTS or performative HTS, and later called Mage in version 2. This implementation can introduce modifications to the speech parameters and their trajectories before they are synthesised with a delay of only one phonetic label. This is achieved by reducing the phonetic context both in training and during synthesis, dropping future dependencies of the current phonetic labels, and creating a buffer that can store the phonetic labels and their parameters. Using this approach an articulation degree real-time interpolation method was evaluated that showed how real time generated trajectories were successfully achieved while maintaining segmental quality. Still, this method needs to be improved, as the substantial drop in context size has an impact on the overall quality of the output.

In the following sections we introduce the systems we used and the experimental design for a listening evaluation using an oracle reactive system against two non-reactive systems, followed by the results of this evaluation. We conclude the report with a discussion of the results and our thoughts on the implications of this investigation.

2. Methodology

Our main hypothesis is that a speech synthesis system that is able to reproduce the modulation between neutral and Lombard speech in different noise conditions will have the intelligibility gains of Lombard speech in noise, while not being penalised for naturalness in quiet conditions. We use an oracle system rather than a live reactive system as this allows us to control the experiment in more detail. The oracle system was designed to simulate a fully reactive system, but to give us more precise control over where the quality of the speech changes. It also provides us with the chance to use the same high quality acoustic models for the synthetic voices that have been evaluated both in stable and noisy environments in [4].

3. Experimental design

We designed an experiment in the train announcement domain, to provide a reason for the noise level to change. In a real train station, the background noise level is stable and generally quiet, but it can increase and decrease sharply as a train passes through the station or starts to leave a station. This peculiar noise shape can happen during announcements and it can mask beyond recognition important pieces of information. We initially used actual recordings of trains, but eventually decided to use speech shaped noise to more accurately evaluate signal-to-noise level consistently and to use a noise type similar to that one used during the training of the non-reactive Lombard voice used in this experiment.

The stimuli were designed using four templates of around 20-25 words from the train announcement domain. Our goal was to reproduce a scenario that the listener could relate to as being akin to being in a train station, where sudden noise can prevent the listener from perceiving key informational items clearly. The sentences included three randomised non-semantically predictable elements (time of departure, destination and platform number) one of which was masked by noise. We selected to mask only the time of departure instead of masking the different elements alternatively to reduce variability even further. Also, the transcription of numbers makes the test more phonetically and semantically balanced and reduces the potential variability in orthography.

The length of the sentences selected had to be short enough to reduce the cognitive load during the transcription task and produce meaningful results. At the same time, it also needed to be long enough to be only partly masked by noise and for a sustained period of time. Different sentences were tested during a pilot test to make sure the difficulty was properly adjusted. The listening test included a transcription task to test intelligibility and a subjective mean opinion score evaluation of naturalness, suitability to the noise conditions, and overall quality.

The two non-reactive voices used for this experiment are the high quality neutral and Lombard voices that were reported in [4]. Both were created using HTS and adapted from a high quality average voice model. The neutral voice was adapted using 2803 Harvard sentences recorded by a British male speaker. The Lombard voice was further adapted from the neutral voice using 780 Harvard sentences recorded while speech shaped noise was being played through headphones at 84 dBA. More information about the creation of the voices can be found in [4]. The batch of randomly generated sentences from the four templates was synthesised using HTS from the two voices.

The oracle system was created using synthesised sentences from both models and semi-automatically replacing the audio segments that contained the time of departure from the neutral sentence with the same segment from the Lombard sentence. The replaced item was deliberately designed to be around the middle of the sentence to match the area where we intended to introduce the masking. The type of noise that we used was female speech shaped noise at 48KHz from the materials provided in the Hurricane Challenge [11]. This noise was also the same one used in the evaluation of the two non-reactive voices in [4] and allowed us to compare directly to that study.

The intelligibility and subjective perception of the quality features of synthesised speech changes with the level and type of noise, as previous evaluations of synthesised Lombard speech have shown [3, 4, 14]. In most cases, evaluation is performed on one or more noise types at three discriminating levels of noise considered to be low, medium and high. The specific

ratios used for these three levels cannot be transferred from one study to the next, because there is a close dependency between the type and level of the noise used, and therefore the efficacy of the masking, and the original levels of the audio materials. In the initial pilot, multiple levels of signal-to-noise ratios (SNR) were tested in stimuli generated with a scaling factor that accounted for the relative levels of speech and noise and rescaling noise accordingly. The scaling factor was defined as:

$$sf(SNR) = \sqrt{10^{\frac{SNR}{10}} \times \frac{P_{noise}}{P_{signal}}} \quad (1)$$

Using the results from the pilot experiment three levels were chosen: -2dB, -6dB and -10dB. A noise track was designed to introduce noise in the middle of each sentence to avoid variability in the relative order of the masked/unmasked segments. Noise was sustained for 2-3 seconds with an additional 1-2 second fade in/out for a more natural effect. Multiple distributions of noise were avoided to reduce the number of tested parameters. The templates were designed with this central position in mind to approximate an ideal length to mask only the selected element. The combination of the sentences and noisy tracks was performed automatically and further corrected by hand. A good alignment was critical for the reactive oracle to avoid the undesired effect of a timing mismatch between the modulated segment and the noise. Using four different announcement templates also introduced variability and balance to prevent semantic predictability.

For the presentation of the experiment we chose a Latin square block-design with 27 participants. All participants were native English speakers with no known hearing issues and were paid for their participation. We arranged them in 9 groups (the combination of the three different systems in three different noise conditions). Each group listened to 45 sentences that included 5 sentences from each of the nine combinations in different orders. The experiment was held in isolated booths and lasted for 45-50 minutes. Stimuli were presented through headphones with fixed volume. During the transcription task participants were allowed to play the recording only once and were asked to transcribe only the three randomised informational elements mentioned before. In the second part of the experiment, participants could play the recording multiple times and needed to provide a score in a 5 point scale according to each of the three dimensions tested: naturalness, suitability and overall quality.

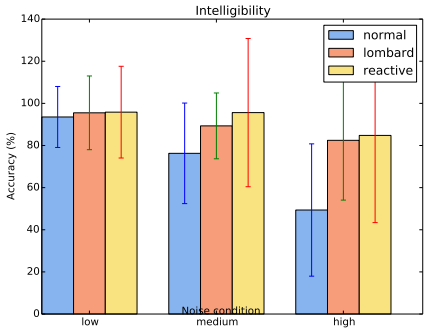


Figure 1: Word accuracy of the elements masked by noise in the three different levels of noise (low: -2dB mid: -6dB and high: -10dB.)

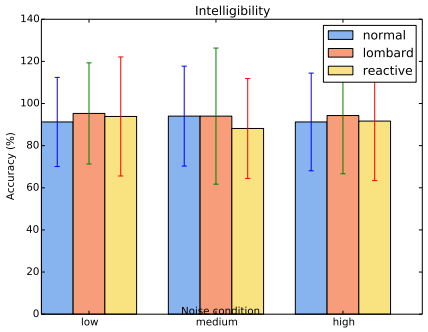


Figure 2: Word accuracy of the noise free elements where part of the utterance is masked by three levels of noise (low: -2dB mid: -6dB and high: -10dB)

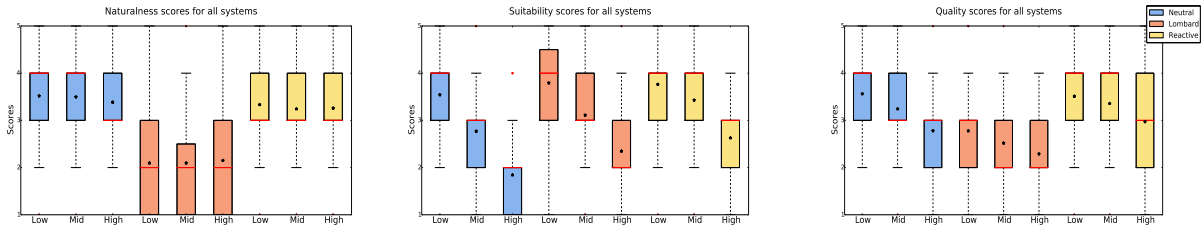
4. Results

Figures 1 and 2 show the intelligibility results from the transcription task. Figure 1 shows the accuracy percentage of the masked element for the three systems in the three different noise conditions. The word accuracy is calculated here only the time of departure item (masked), and by aggregating the results for each of the four digits of the time for each response, as many participants made partial guesses. Figure 2 shows the results of the unmasked elements in each utterance, departure and platform number.

Figure 1 shows that the reactive oracle system is generally as intelligible as the Lombard system in all of the noise conditions when the masking is present, while the neutral system drops the accuracy much more (from 94% to 65%) as the noise level is increased. Figure 2 shows that the overall intelligibility of the unmasked components remains reasonably constant according to the voice used and the noise level elsewhere in the utterance.

Analysis of Variance showed that for elements in the unmasked condition there is no significant effect for either the voice type or noise condition. However, for elements in the masked condition there is an effect for both parameters. To test between which categories of voice and noise these differences lie, Post-hoc Bonferroni corrected t-tests were performed. We found significant differences ($t = 4.50, p < 0.01$ and $t = 5.97$, $p < 0.01$)) in intelligibility between noise conditions low and mid, and mid and high, and significant differences between the normal voice and the other two voices ($t = -6.59$, $p < 0.01$ and $t = -7.9, p < 0.01$), but no difference between the intelligibility of the lombard voice and the reactive voice ($t = -1.57$, $p = 1.16$). We therefore confirm our hypothesis that the reactive system does not suffer an intelligibility decrease compared to the neutral voice as intelligibility is increased by using Lombard speech whilst in noise.

Figure 3 presents the results for the analysis of the mean opinion scores for the three evaluated dimensions in standard boxplots. The red line represents the median and the star sign inside each box represents the average value. The edges of the box represent the Q1 and Q3 quartiles and the whiskers extend 1.5 the interquartile range. Under each box are the non-parametric statistical tests to compare the mean ranks of the scores using a Wilcoxon signed-rank test. The cells marked with an X in the table represent pairs of systems where there was a significant statistical difference at $p < 0.01$.



	Neutral low	Neutral mid	Neutral high	Lombard low	Lombard mid	Lombard high	Reactive low	Reactive mid	Reactive high
Neutral-low				x	x	x			
Neutral-mid				x	x	x			
Neutral-high				x	x	x			
Lombard-low	x	x	x				x	x	x
Lombard-mid	x	x	x				x	x	
Lombard-high	x	x	x				x	x	x
Reactive-low				x	x	x			
Reactive-mid				x	x	x			
Reactive-high				x		x			

	Neutral low	Neutral mid	Neutral high	Lombard low	Lombard mid	Lombard high	Reactive low	Reactive mid	Reactive high
Neutral-low			x			x			x
Neutral-mid									
Neutral-high	x			x	x		x	x	
Lombard-low			x			x			x
Lombard-mid			x						
Lombard-high	x			x			x		
Reactive-low			x			x			x
Reactive-mid			x						
Reactive-high	x			x			x		

	Neutral low	Neutral mid	Neutral high	Lombard low	Lombard mid	Lombard high	Reactive low	Reactive mid	Reactive high
Neutral-low					x	x			
Neutral-mid									
Neutral-high									
Lombard-low									
Lombard-mid	x								
Lombard-high	x						x		
Reactive-low						x			
Reactive-mid									
Reactive-high									

Figure 3: Boxplot results of the mean opinion subjective scores for naturalness, suitability and overall quality. The inferential statistic results for each dimension are presented under each graph. The pair marked with an X represents a significant difference between pair of systems using a Wilcoxon signed paired test at $p < 0.01$.

The leftmost panel shows that when subjects were asked to judge naturalness of the voices, they judged the neutral and reactive voices as more natural than the Lombard voice even in the least noisy conditions. This is consistent with our initial hypothesis. The oracle system was perceived to be significantly more natural than the Lombard system and as natural as the neutral system in all noise conditions.

The results for suitability, displayed in the central panel, show that the perceived suitability of the neutral system decreases as the level of noise increases and is statistically significant between the high and low noise conditions. The same results can be seen for the Lombard voice and the the reactive voice, although the absolute drop in suitability is less for each respectively as the noise level increases. There is a trend for the reactive voice to be considered more suitable for the mid and high noise conditions, but there are no actual statistical significances here.

The far right panel shows the results for overall quality, where very few differences are seen. This result is expected as evaluating overall quality is far more subjective as it is a complex feature that each participant can consider differently, which can introduce variability and less clear preference between systems.

5. Discussion

The results presented above show how a noise-reactive TTS oracle is able to display significant improvements in intelligibility compared to non-reactive systems without loss of naturalness. These results have been achieved in very controlled noise conditions with one type of noise and three levels of SNR. Given the idiosyncratic characteristics of the Lombard modulation, it would be interesting to take this experiments further by creating a fully functional reactive system and making more thorough tests with other types of noise and noise distributions.

It would also be necessary to make certain design decisions for a full implementation. The most salient would be, first, how to extract and process the noise input as a control variable for the system. Second, the selection of activation and control parameters, which would include the definition of thresholds, the shape of the trajectory of the modulation, either linear or stepwise, and, in the case of a linear approach, the selection of an architecture that can allow interpolated adaptation. Lastly, the effect of latency would have to be assessed to see to what extent it is detrimental.

6. Conclusions

In this investigation we have proposed a speech synthesis system reactive to dynamic noise conditions using neutral and Lombard speech. We hypothesised that such a system would be able to combine the advantages of both styles and avoid the disadvantages. The results of our listening experiment have obtained statistically significant gains for intelligibility and naturalness from a proposed oracle reactive system compared to both neutral and Lombard non-reactive systems. The intelligibility test that we performed showed how the system was able to perform as good as the Lombard system in the noisy segments in all noise conditions. The subjective evaluation showed that the oracle reactive system was perceived to be as natural as the neutral system, while the Lombard system scored significantly lower. These results are encouraging for reactivity as they provide with the first intelligibility evaluation focused on modulated synthesised speech under changing noise with a subjective evaluation of naturalness, suitability and quality. They show that reactivity can be an interesting implementation feature that can have the potential to improve performance in changing noise against non-reactive systems. The creation of the stimuli using an oracle has also provided us with a chance to highlight some necessary precautions that would have to be accounted for if presented with the chance to build a reactive system.

7. Acknowledgements

The authors would like to thank Cassia Valentini-Botinhao for giving access to the two TTS voices used during this investigation. This work is partly supported from the European Community's Seventh Framework Programme (F P7/2007-2013) under grant agreement 287678 (Simple[4]All)

8. References

[1] Langner, B. and Black, A., "Improving the understandability of speech synthesis by modeling speech in noise", Proc. ICASSP, 1:265-268, 2005.

[2] Venkatagiri, H., "Segmental intelligibility of four currently used text-to-speech synthesis methods", The Journal of the Acoustical Society of America, 133:2095, 2003.

[3] Raitio, T., Suni, A., Vainio, M. and Alku, P., "Analysis of HMM-Based Lombard Speech Synthesis", INTERSPEECH, 2781-2784, 2011.

[4] Valentini-Botinhao, C., Yamagishi, J., King, S., "Evaluating speech intelligibility enhancement for HMM-based synthetic speech in noise", Proc. Sapa Workshop, 2013.

[5] Lombard, E., "Le signe de l'elevation de la voix", Ann. Maladies Oreille, Larynx, Nez, Pharynx, 37(101-119):25, 1911.

[6] Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. and Stokesm M. , "Effects of noise on speech production: acoustic and perceptual analyses", The Journal of the Acoustical Society of America, 84(3):9-17, 1988.

[7] Junqua, J. C., "The influence of acoustics on speech production: a noise induced stress phenomenon known as the Lombard reflex", Speech Communication, 20(1):13.22, 1996.

[8] Cooke, M., King, S., Garnier, M., Aubanel, V., "The listening talker: a review of human and algorithmic context-induced modifications of speech". (in press) Computer, Speech and Language.

[9] Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y., "Evaluating the intelligibility benefit of speech modifications in known noise conditions", Speech Communication, 55:572-585, 2013.

[10] King, S., Vasilis, K., "The Blizzard Challenge 2010", The Blizzard Challenge Workshop.2010.

[11] Cooke, M., Mayo, C., Valentini-Botinhao, C., "Intelligibility enhancing speech modifications: the Hurricane Challenge". Proc. Interspeech, Lyon, France, 2013.

[12] Cernak, M., "Unit selection speech synthesis in noise", IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, 2006.

[13] Zen, H., Tokuda, K., Black, A., "Statistical parametric speech synthesis", Speech Communication, 51(11):1039-1064, 2009.

[14] Raitio, T., Suni, A., Vainio, M., Alku, P., "Synthesis and perception of Breathy, Normal, and Lombard speech in the presence of noise", Computer Speech and Language, 2013.

[15] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm", IEEE Transactions on Audio, Speech and Language Processing, 17(1): 66-83, 2009.

[16] Valentini-Botinhao, C., Yamagishi, J., King, S., "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?", Interspeech, 1837-1840, 2011.

[17] Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., Zen, H., "Cepstral analysis based on the Glimpse Proportion measure for improving the intelligibility of HMM-based synthetic speech in noise", ICASSP, 3997-4000, IEEE, 2012.

[18] Moore, R., "Presence: a human-inspired architecture for speech-based human-machine interaction", IEEE Transactions on Computers, 56(9):1176-1188, 2007.

[19] Moore, R., Nicolao, M., "Reactive speech synthesis: actively managing phontic contrast along an H&H contnuum", 7th Internat. Cong. on Phonetic Sciences, pages 1422-1425, 2011.

[20] Dutoit, T., Astrinaki, M., Babacan, O., d'Alessandro, N., Picart, B., "pHTS for Max/MSP: a streaming architecture for statistical parametric speech synthesis", Technical report, 2011.

[21] Astrinaki, M., d'Alessandro, N., Picart, B., Drugman, T., Dutoit, T., "Reactive speech synthesis and continuous control of HMM-based speech synthesis", Spoken Language Technology Workshop, IEEE, 252-257, 2012.