Using neighbourhood density and selective SNR boosting to increase the intelligibility of synthetic speech in noise

Using neighbourhood density and selective SNR boosting to increase the intelligibility of synthetic speech in noise

Cassia Valentini-Botinhao, Mirjam Wester, Junichi Yamagishi, Simon King

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

C.Valentini-Botinhao@sms.ed.ac.uk, mwester@inf.ed.ac.uk

jyamagis@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

Motivated by the fact that words are not equally confusable, we explore 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 to the detriment of another word. Sentence intelligibility did not generally 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, 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 our results indicate – semantic, syntactic and acoustic information about the word and the sentence.

Index Terms: word confusability, neighbourhood density, HMM-based speech synthesis

1. Introduction

Text-to-Speech (TTS) systems are deaf (and blind) to the environment and currently do not react to adverse conditions when intelligibility is possibly more important than naturalness. Although research aimed at generating clear or Lombard style synthetic speech has been carried out [1–5] such speech modification methods do not take into account word-level confusability. The best objective measures of intelligibility [6–8] are based on the acoustic and effective signal processing that takes place in the human auditory system and does not consider lexical activation or any further context information.

However, some words are inherently more intelligible than others i.e., they are less likely to be confused with other words. This property of words is currently ignored but could potentially be exploited when applying speech modifications. The premise is that modifications aimed at improving intelligibility should not necessarily be "on" the whole time. For example, we can think in terms of energy budget whereby energy in a sentence is reallocated on the basis of the expected intelligibility of a word. More or less energy is expended depending on the predicted intelligibility of a word. Another approach would be to use word confusability to control the balance between naturalness/quality of the speech and any intelligibility improvements resulting from the modification. In this case, the level of modification could be constrained by the degree of distortion it introduces. As it is not clear how to define what an acceptable amount of distortion is, we use word-level information to reallocate energy under the constraint of fixed sentence energy.

This work is a first attempt towards making use of a model of spoken word activation, the neighbourhood activation model (NAM) [9], in an energy-based speech modification. We address two questions: can neighbourhood density values be used to predict intelligibility at the word-level and can we use this information to improve overall word recognition by selectively boosting highly confusable words.

Of course there are many factors that influence the intelligibility of a word: acoustic confusability, linguistic confusability, the inherent intelligibility of a speaker, environmental factors (e.g., noise types) and listener characteristics. How to predict which words in a sentence are going to be easily intelligible and which ones hard is therefore not straightforward. Furthermore, in order to measure the effectiveness of selectively boosting words based on their intelligibility, the influence of all these different factors needs to be restricted. Therefore, we decided on the following constraints in our experiments. We only consider confusability at the word-level (no linguistic confusability), synthetic speech from one speaker and one type of noise (speech-shaped noise). The modification we are looking at is energy reallocation, which we view as a starting point for other types of modifications.

The remainder of this paper describes three listening tests. In the first experiment, we investigated the use of neighbourhood density as a predictor of word intelligibility of synthetic speech in noise for words in isolation. In the second and third experiments, the words from the first experiment were placed in matrix style sentences and energy reallocation was applied aiming for maximum intelligibility in two different ways. The following sections describe the set up of the experiments, our findings and a discussion of the results.

2. Methodology

2.1. Word selection criteria

To test different word boosting strategies it is important to select words that cover a wide range of acoustic confusability, that is easy- and hard-to-understand words. One way of performing this categorization is to use the neighbourhood density value of words. Lexical or phonological neighbourhood density (ND) plays an important role in word recognition. Words with many lexical neighbours, differing by one phoneme insertion, deletion or substitution are more difficult to recognise that words with few lexical neighbours [9]. In De Cara and Goswami [10] a second definition of phonological neighbourhood is given: the OVC-metric. In this metric, words that differ by insertions,

Cassia Valentini-Botinhao, Mirjam Wester, Junichi Yamagishi, and Simon King. Using neighbourhood density and selective SNR boosting to increase the intelligibility of synthetic speech in noise. In 8th ISCA Workshop on Speech Synthesis, pages 133-138, Barcelona, Spain, August 2013.

deletions or substitutions in either the onset, vowel or coda of a word are counted. According to the OVC-metric not only words like *main* and *gain* are phonological neighbours but also for example *main* and *strain*.

In our study, we set out to define a set of "hard" words and a set of "easy" words in terms of intelligibility. The words were selected to fill slots in Matrix-style sentences [11] of the form: [imperative verb] the [adjective] [adjective] [noun]. We chose 10 verbs, 20 adjectives and 10 nouns from an existing monosyllabic lexical database which contained both neighbourhood density statistics and frequency statistics [10]. Our criteria, similar to those used by [12], were:

- written and spoken frequency $\geq$ 10 per million,
- per Matrix slot:
  - 5 "hard" words, i.e., from a dense neighbourhood, ND-OVC $\geq$ 37,
  - 5 "easy" words, i.e., from a sparse neighbourhood, ND-OVC $\leq$ 17.

The intervals of ND-OVC values that define the easy and hard categories were as far apart as possible under the word frequency constraints.

2.2. Synthetic speech and noise material

To build the HMM-based TTS voice for this work we used read speech recordings of a British male speaker. The voice was created from a high quality average voice model which was adapted to the speaker's voice using three hours of his speech sampled at 48 kHz as described in [5]. We used a hidden semi-Markov model as the acoustic model. The isolated words were synthesised in a carrier sentence of the format: "Now we will say "pause" *word* "pause" again". They were then automatically segmented and added to noise with 200 ms initial and final lags. The speech-shaped noise was generated using recordings of a female speaker sampled at 48 kHz (similar to [13]). Different signal-to-noise ratio (SNR) values were obtained by varying the level of the speech stimuli against a constant level of speech-shaped noise (similar to [9]).

2.3. Procedure

Stimuli were presented to native British English speakers with no hearing problems over Beyerdynamic DT770 headphones in individual sound-treated booths. The experiments were run using a custom-built MATLAB software application. Each stimulus was presented once. Listeners typed what they heard, after which the following stimulus was presented. The listeners were instructed to type 'X' if they could not make out the word(s). Word accuracy rate (WAR) was calculated as the percentage of correct word transcriptions across the listeners. Homophones, for example, a response of "sea" for "see" were considered correct.

3. Listening tests and results

First a listening test of words in isolation is described. The goal of this experiment is to find the "true" intelligibility scores of words, rather than the intelligibility expectation based only on the ND values of the words. (Although one would expect them to be similar). This is followed by sentence experiments in which the goal is to improve overall sentence intelligibility by using this prior information about word intelligibility. To this end, sentence experiments using two energy reallocation strategies were investigated. In the first one, energy is taken from one word and given to another: the giver/receiver strategy. The second strategy involves reallocating energy from the whole sentence to boost one word.

3.1. Isolated word experiment design

To find which words can benefit from being presented at higher SNR levels we need to obtain word intelligibility scores at a range of different SNR values. Before the actual listening test could be performed we needed to find the range of SNR values at which to present the isolated words. The range needed to be such that "hard" words were intelligible at the highest SNR level and "easy" words unintelligible at the lowest level. A separate listening experiment involving 10 participants was carried out to find the range. On the basis of their results five SNR values were chosen: -8, -3.5, 1, 5.5 and 10 dB.

The 40 words were presented at each of the five SNR levels (200 stimuli) randomised over four blocks (50 words per block). In each block, the SNR values were ordered from low to high. Prior to the main test, listeners received a practise session presented at a mid range SNR, using 20 words from outside the test set. At the end of the test, the participants were asked to transcribe the words in clean condition, i.e. no speech-shaped noise present. 25 listeners performed the isolated word task.

3.2. Isolated word results

Figure 1 shows scatter plots of the WAR results obtained for different ND values in clean (top) and in noise at a SNR = 5.5 dB (bottom). The results show that even in the clean condition a number of words were poorly understood, achieving less than 60% WAR. Most of these words are from a dense neighbourhood, belonging to the "hard" category. Although the linear correlation between ND and WAR is quite low (-0.46 for the clean condition and -0.31 for the noisy condition) when comparing the scatter plots of the clean and noisy conditions we can see that the "easy" words are more robust to noise. That is, the dispersion towards the low WAR region caused by the presence of noise is smaller.

As each word was presented in noise at five different SNRs we are able to draw psychometric curves for each individual word. We present the curves for verbs in Figure 2. (The results for nouns and adjectives are similar but are not presented here for brevity's sake). According to the NAM model we expect words classified as "easy" to have higher WAR than "hard" words. We can see however some words do not behave as expected. For instance the verb *have* classified as an easy word is in fact less intelligible than expected and that the verb *see* is easier to recognise than expected from its ND value. This mismatch between ND values and intelligibility scores of synthetic speech in noise is not wholly unexpected as the ND does not account for noise and type of speech (TTS).

To illustrate how challenging it is to "represent" intelligibility at the word-level we use the glimpse proportion measure (GP) [6] as a reference. We calculate the GP for each of the words in the five different SNR conditions and correlate that with the subjective scores. The GP measure was shown to obtain a high correlation coefficient (up to 0.94) with subjective intelligibility scores of a male TTS voice in diverse noise conditions when both GP and WAR scores were calculated at a word-level but averaged across the different words [14]. Here the GP values however a very poorly correlated (0.44) to WAR scores calculated for individual words.

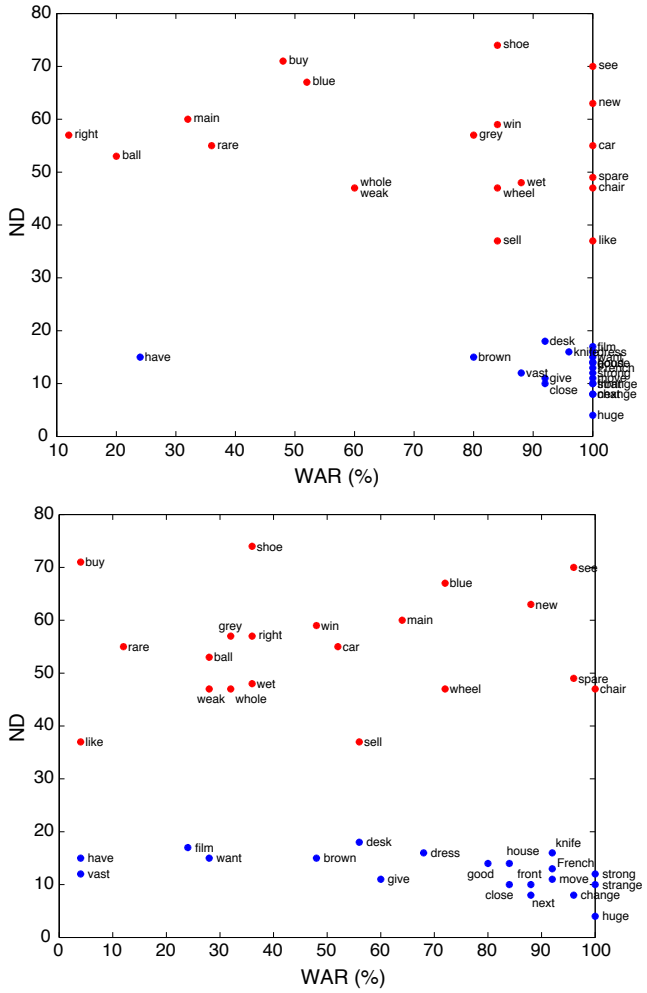


Figure 1: Neighbourhood density (ND) versus word accuracy rates (WAR) in clean (top) and in noise at a SNR = 5.5 dB (bottom), red and blue dots represent hard and easy words.

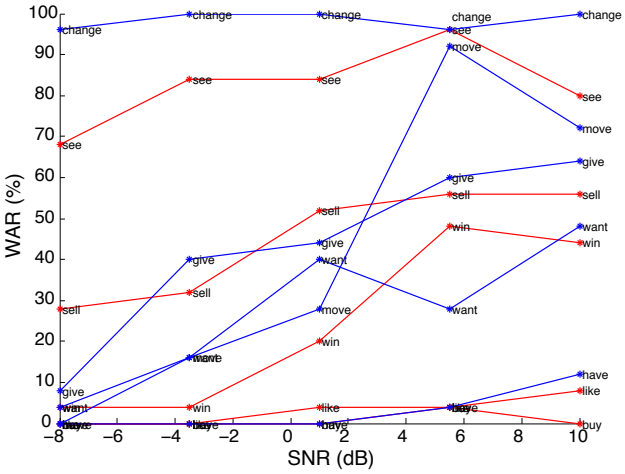


Figure 2: Psychometric curves for verbs, red and blue lines represent hard and easy words.

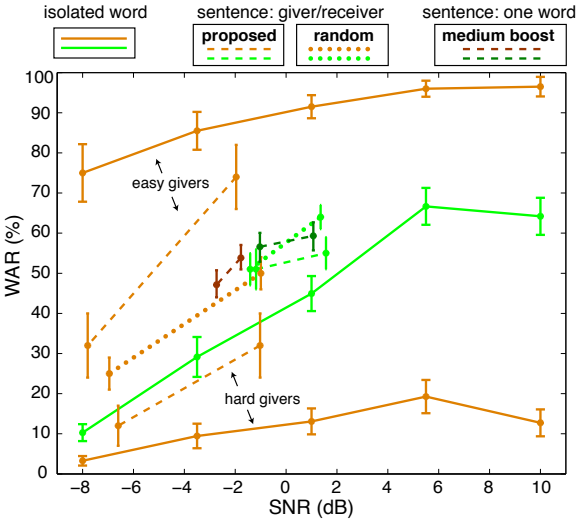


Figure 3: Psychometric curves for givers (oranges) and receivers (greens) in isolation and in sentences. "Givers" are divided into easy and hard for isolated and proposed conditions.

3.3. Sentence experiment: Giver/receiver boosting

The 40 words from the isolated word experiment were split into two categories: givers and receivers. The reallocation strategy used here is that energy is taken from one word (the giver) and given to another word (the receiver), keeping the overall energy budget the same. A word was considered a giver if the WAR in isolation for all SNRs tested as either quite low – hard giver – or quite high – easy giver. Easy and hard now relate to the words' intelligibility scores rather than their ND values. The expectation is that easy givers are robust and attenuating them will not harm their intelligibility much, whereas hard givers will remain unintelligible no matter what so they are not worth spending energy on. The receivers were words that showed steep slopes in the isolation experiment, providing evidence that at higher SNR values they were more intelligible. Our expectation is that they would benefit from energy boosting. Figure 3 shows psychometric curves for all listening tests described in this paper. Here we focus on the curves for "isolated words" – solid lines – obtained by averaging WAR values across receivers and givers. Both giver curves (easy and hard orange solid lines) are quite flat across the SNR range and on average receivers (solid green line) are more sensitive to changes in SNR, which makes them promising candidates for selective boosting.

3.3.1. *Sentence material*

To create the sentence material we built Matrix-style sentences of the form: [imperative verb] the [adjective] [adjective] [noun], for example: "Change the grey strange dress". Each sentence contains one giver (of energy) word and one receiver (word). The other words in the sentence (fillers) were randomly selected from the remainder of the 40 word pool, each word occurred six times. Varying the position of giver and receiver in the sentence results in 12 possible sentence types. Per sentence type, five sentences were created, resulting in 60 sentences in total. These sentences contain very little context information, aside from the structure there is no further linguistic information available to listeners, i.e., the words are equally predictable (or unpredictable).

3.3.2. Modifications

To investigate whether selectively boosting the receiver word by attenuating the giver word increases overall intelligibility we compared two types of modifications:

- **proposed** - select giver/receiver pairs according to the results of the isolated word experiment;
- **random** - select giver/receiver pairs randomly.

From the results of the isolated word experiment we expect that different receivers need different amounts of boosting to raise their WAR to a similar value, but to simplify the experiment we fix the amount of power level loss to 6 dB. On average the receivers' power increases by 2.7 dB with the constraint that the overall energy of the sentence remains unchanged.

3.3.3. Listening experiment

As words in isolation require higher SNRs to be intelligible than words in a sentence we carried out a pre-test (five participants) to find the SNR level (-3 dB) that resulted in an average of 50 % WAR across all selected sentences.

All 60 sentences were evaluated for the three different conditions: the two modifications and unmodified. As we did not want listeners to hear a sentence more than once the experiment was divided across three groups of listeners. Each listener heard all 60 sentences once and the modification type applied to each sentences was spread across the listeners so the whole test (180 stimuli) was covered by three listeners. Prior to the main test participants carried out a practice session consisting of 20 sentences of the same structure as the test however filled with other words. At the end of the test all participants were also asked to transcribe the 60 sentences in clean condition. In total, 60 listeners performed the test.

3.4. Giver/receiver sentence results

We present the average WAR of easy and hard words in Table 1 obtained in isolation and in a sentence. Easy words are more intelligible in both scenarios but the difference is less pronounced in a sentence than in isolation. These results indicate that the effect of neighbourhood density on the intelligibility of words is limited when a word is presented with context.

	easy words	hard words
isolation	93.2 (3.8)	71.2 (6.5)
sentence	97.8 (0.8)	94.4 (2.0)

Table 1: Mean word recognition (%) and its standard error for easy and hard words in isolation and in a sentence in clean conditions.

Figure 4 gives the results averaged across words and listeners for each modification in terms of absolute change compared to the unmodified case. As a reference, the rates obtained for unmodified speech were: WAR = 49.6% and for proposed/random: WAR_R = 51.25/53.3%, WAR_G = 50.5/49.9% and WAR_F = 48.3/47.6%. ($_\text{R}$ = receivers, $_\text{G}$ = givers, $_\text{F}$ = fillers). We can see that boosting a word at the detriment of another word decreases WAR results for both modifications. The intelligibility of the givers drops significantly in both cases, more than a 25% absolute drop, while the receivers only gain up to 12% in word accuracy. The results also show that on average choosing the pairs randomly rather than according to the isolated word experiment generates a larger gain for receivers

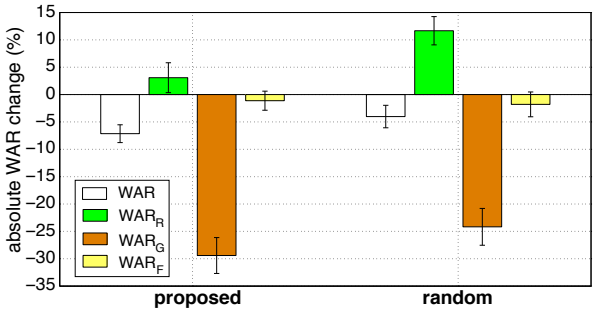


Figure 4: Absolute changes in WAR (in %) of proposed and random modifications with respect to unmodified sentences. Sentence SNR=-3 dB. $_\text{R}$ = receivers, $_\text{G}$ = givers, $_\text{F}$ = fillers.

and a smaller WAR drop for givers. A sentence-based analysis showed that the intelligibility of filler words changed significantly in some contexts even though no modification was applied to these words.

To compare the results across the two experiments, the SNR that each giver and receiver word was presented at in the sentence experiment is calculated. As the same word appears as a giver (or receiver) in more than one sentence, we obtain the word's SNR by averaging across its occurrences (as either a giver or receiver, fillers are not included here). This gives us a unique SNR per word. The results are then averaged within each category: easy givers, hard givers and receivers. These values and their standard error (which represents the variance across the words within a category) are shown in Figure 3 (sentence:giver/receiver curves) in addition to the earlier discussed results for words in isolation. Note that the sentence results only contain two points along the x-axis (SNR), because words were either boosted (receivers) or attenuated (givers), whereas in the isolated word experiment words were played at five different SNR values.

Looking at the proposed modification (dashed line) it can be seen that the hard givers and receivers are more intelligible in a sentence than in isolation (WARs: 11% to 31% hard giver; 31% to 58% receiver) whereas the easy givers are on average less intelligible in a sentence (WARs: 88% to 74%). The slopes of the curves are also different, that is, the easy and hard givers' WAR drops more than expected and receivers' WAR does not increase as much. It seems too much energy is taken from the givers while the receivers are not getting enough, which explains why the WAR per sentence does not increase.

The dotted line in Figure 3 shows the psychometric curves for randomly chosen receivers and givers (only one curve as there is no notion of easy and hard givers in the random condition). Choosing giver and receiver pairs randomly brings their psychometric curves closer to each other. Although the increase in SNR value is similar across proposed and random modifications, the intelligibility of words in the random selection improves more. This seems to be caused by the fact that words originally classified as hard givers are now in the receiver category. Basically the hard givers – words for which we expected boosting to be ineffective based on their scores in isolation– benefit most from boosting.

3.5. Sentence experiment: boosting one word

The previous experiment showed us that boosting one word and attenuating another word in the same sentence impacts on the

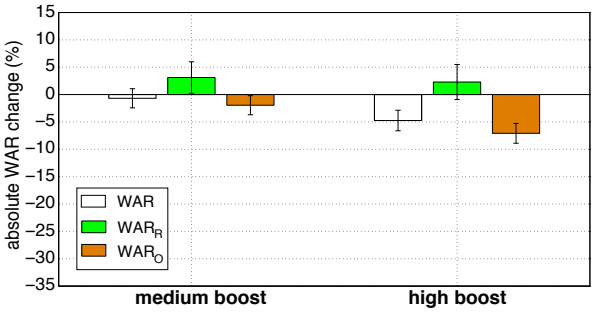


Figure 5: Absolute changes in WAR (in %) of the medium boost and high boost modifications with respect to unmodified sentences. Sentence SNR=−3 dB. R = receivers, O = others.

intelligibility of the other words in the sentence. Not only that, taking energy from one word to give to another does not improve overall intelligibility rates, mostly because the intelligibility of the attenuated word drops at a much higher rate than that the intelligibility of the boosted word increases. To overcome these two issues, we use a different type of energy reallocation strategy: energy is reallocated from the whole sentence to boost just one word. This can be viewed as emphasising a word in a sentence while making the rest of the words more quiet, possibly a slightly more natural occurring modification. Even though we saw, in the previous experiment, that randomly selecting givers and receivers resulted in higher receiver gains we have kept the same set of receiver words in this third experiment to be able to analyse the proposed selection under a more promising modification strategy.

3.5.1. Modifications

To investigate whether boosting one word in the sentence while keeping the overall SNR fixed (−3 dB) increases intelligibility we evaluate the following modifications:

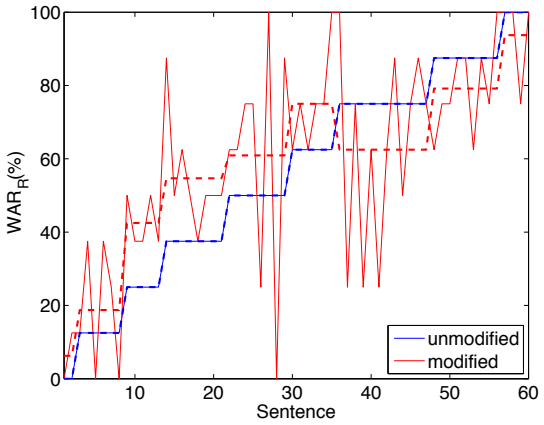
- **medium boost** - boost the receiver word by 3 dB and attenuate sentence;
- **high boost** - boost the receiver word by 5 dB and attenuate sentence.

3.5.2. Listening experiment

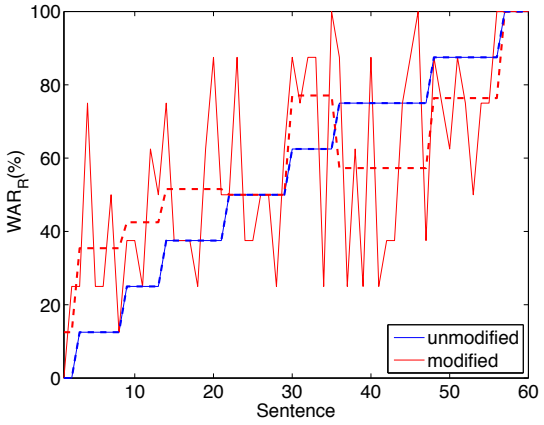
The set-up of this experiment (boosting a single word) was exactly like the giver/receiver listening test. 24 participants took part in this experiment.

3.6. Results for boosting one word experiment

Figure 5 shows WAR results averaged across words and listeners in terms of absolute change compared to the unmodified case for medium and high boost modifications. WAR_O refers to the intelligibility of the other words in the sentence –the attenuated words. As a reference, the WAR values obtained for unmodified speech were: WAR = 54.5 %, WAR_R = 56.0 % and WAR_O = 53.9 %. Although there is still an overall drop in intelligibility the drop is much smaller than what was observed in the previous experiment, see Figure 4. Particularly, we note that the medium boost modification WAR_R gains are comparable to the ones obtained using the proposed modification in the previous experiment (around 3.0 % absolute gain). However, the loss in WAR_O is much smaller (from 29.0 % to 1.9 % absolute drop)



(a) Medium boost



(b) High boost

Figure 6: Receivers' word accuracy rate (WAR_R) for each of the 60 sentences. The sentence index is ordered according to the unmodified scores (blue). Modified scores are presented at the sentence level (red continuous line) and the sentence interval level (red dashed line). Sentence SNR=−3 dB.

indicating that boosting one word in a sentence is a much better strategy.

Figure 3 shows the psychometric curves for receivers and others in the medium boosting condition (sentence: one word curves). We can see that the slope for receivers (dark green dashed line) for the medium boost modification is similar to the slope for the proposed modification (light green dashed line) and that the slope for others (i.e. givers, dark orange dashed line) is similar to the easy givers (light orange dashed line).

To identify under which conditions there was an increase of word intelligibility Figure 6 presents a sentence level analysis: WAR_R results for each of the 60 sentences. The sentences are ordered according to the WAR_R obtained in the unmodified case to check how this affect the results. The continuous red curve represents the WAR_R for the medium boost (top) and the high boost (bottom) modifications. The dashed red curves represent these results averaged across each sentence interval. A sentence interval is taken as the range where unmodified WAR_R results are constant. It can be seen that for highly intelligible words boosting can decrease the WAR_R, for both medium and high boost modifications. It seems that if a word is more intelligible than a certain threshold boosting is harmful and that this thresh-

old depends on the level of boosting. We can also see that the effect of the boosting value depends on the WAR of the receiver: poor receivers should be boosted more and highly intelligible receivers should not be boosted at all. The best strategy is then to boost the most unintelligible words in the sentence and apply an energy boost inversely proportional to the intelligibility of the word.

4. Discussion and Conclusions

This study aimed to capitalise on the idea that modifications should not be "on" the whole time but rather applied more judiciously to enhance intelligibility of synthetic speech in noise. Our experiments were designed to constrain the factors that influence the predictability of words to acoustic level confusability, to enable us to measure the effectiveness of boosting words based on their intelligibility. We carried out an isolated word experiment with 20 "hard" words from a dense and 20 "easy" words from a sparse neighbourhood according to the OVC metric in order to cover a wide range of confusability. However, our results showed that not only does neighbourhood density (ND) affect the intelligibility of words in isolation but the type of speech –a TTS voice– , the noise –speech-shaped noise– and the lexical complexity of each word also influence a words' intelligibility. Therefore, instead of using ND to select words to boost and attenuate we used the actual subjective intelligibility scores of words in isolation.

Two sentence experiments were performed using a set of 60 Matrix-style sentences which were created using the 40 easy and hard words. In the first experiment, the modification strategy was to boost one word –the receiver– by 2.7dB while attenuating another –the giver– by 6dB. The results showed that boosting a word to the detriment of another is not a good strategy, independent of the selection of the words: the intelligibility of the giver word drops by 30% absolute while receivers only increase by 3%. Moreover selecting word pairs according to their intelligibility scores in isolation performed worse than selecting them randomly. The psychometric curves for intelligibility of words in speech-shaped noise change significantly when going from isolation to a sentence, both in terms of offset and slope. Intelligibility scores of words in isolation are a poor predictor of intelligibility scores in a sentence. Furthermore, the intelligibility of words whose energy remained unmodified also changed, showing the giver/receiver strategy is not an appropriate strategy.

In the second sentence experiment, the modification strategy was to boost one word while attenuating all the other words in the sentence. The results show that this is a better modification strategy as the decrease in intelligibility for givers went from 30% to only 3%. Spreading the attenuation across all other words in a sentence is beneficial as well as being more natural. The overall gains in the intelligibility of the receiver words was still limited. Analysis at a sentence level showed that boosting is most beneficial when the intelligibility of a word is poor and the boosting level is appropriate. Boosting becomes harmful when the intelligibility of the word is already reasonable to good.

Selectively boosting words by reallocation of energy can be useful for improving intelligibility in noise, but only if the word is poorly understood to start with. If we have reliable ways of predicting word-level intelligibility it is possible to increase sentence intelligibility by selectively boosting the energy of a highly confusable word. This is a promising result that advocates the use of word intelligibility scores as prior knowledge for more complex modifications. The poor word-level intelligibility prediction results using the neighbourhood density and the glimpse proportion measure indicate however that much work needs to be done in order to obtain reliable measures of word-level confusability even for the simplest scenario of words in isolation. Translating that to sentences is an even larger challenge. The fact that subjective scores of words in isolation hardly reflect their scores in a sentence indicates that this prediction has to consider the context of the word in a sentence, not only for the additional linguistic cues but also for the acoustic coarticulation cues as well.

5. Acknowledgement

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213850 (SCALE) and 256230 (LISTA), and from EPSRC grants EP/I031022/1 (NST) and EP/J002526/1 (CAF).

6. References

[1] B. Langner and A. W. 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. K. 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," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

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

[8] C. H. Taal, R. C. 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] P. Luce and D. Pisoni, "Recognizing spoken words: The neighborhood activation model," *Ear and hearing*, vol. 19, no. 1, pp. 1–36, 1998.

[10] B. Cara and U. Goswami, "Similarity relations among spoken words: The special status of rimes in English," *Behavior Research Methods, Instruments and Computers*, vol. 34, pp. 416–423, 2002.

[11] W. A. Dreschler, "Hearing in the communication society D-2-2 deliverable," 2006. [Online]. Available: http://hearcom.eu

[12] M. Cooke, "Discovering consistent word confusions in noise," in *Proc. Interspeech*, Brighton, U.K., 2009.

[13] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Comm.*, vol. submitted, 2012.

[14] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Florence, Italy, August 2011.