Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 572–585

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Evaluating the intelligibility benefit of speech modifications in known noise conditions

Martin Cooke [a,b,*], Catherine Mayo [c], Cassia Valentini-Botinhao [c], Yannis Stylianou [d], Bastian Sauert [e], Yan Tang [b]

[a] Ikerbasque (Basque Science Foundation), Bilbao, Spain
[b] Language and Speech Laboratory, Universidad del Pais Vasco, Vitoria, Spain
[c] Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[d] ICS-FORTH, Institute for Computer Science, Crete, Greece
[e] Institute of Communication Systems and Data Processing, RWTH Aachen University, Aachen, Germany

Received 26 June 2012; received in revised form 14 December 2012; accepted 1 January 2013
Available online 11 January 2013

Abstract

The use of live and recorded speech is widespread in applications where correct message reception is important. Furthermore, the deployment of synthetic speech in such applications is growing. Modifications to natural and synthetic speech have therefore been proposed which aim at improving intelligibility in noise. The current study compares the benefits of speech modification algorithms in a large-scale speech intelligibility evaluation and quantifies the equivalent intensity change, defined as the amount in decibels that unmodified speech would need to be adjusted by in order to achieve the same intelligibility as modified speech. Listeners identified keywords in phonetically-balanced sentences representing ten different types of speech: plain and Lombard speech, five types of modified speech, and three forms of synthetic speech. Sentences were masked by either a stationary or a competing speech masker. Modification methods varied in the manner and degree to which they exploited estimates of the masking noise. The best-performing modifications led to equivalent intensity changes of around 5 dB in moderate and high noise levels for the stationary masker, and 3–4 dB in the presence of competing speech. These gains exceed those produced by Lombard speech. Synthetic speech in noise was always less intelligible than plain natural speech, but modified synthetic speech reduced this deficit by a significant amount.

Keywords: Speech intelligibility; Speech modification; Synthetic speech

1. Introduction

Speech output, whether spoken live, recorded, or generated synthetically from text, is used in a growing range of applications, including public address systems, vehicle navigation devices and mobile phones, and is likely to become more widespread in domestic situations for interaction with consumer devices and speech-based warning systems. Maintaining intelligibility in such settings without resorting to increases in output level is a challenge, particularly in the presence of additive and convolutional distortions. Unlike current speech output technology, human talkers appear to adapt to the immediate context by changing the acoustic, phonetic, and linguistic content of their speech (Lindblom, 1990; Picheny et al., 1985; Summers et al., 1988; Howell et al., 2006; Uther et al., 2007; Patel and Schell, 2008; Cooke and Lu, 2010). Recently, a number of speech modification algorithms designed to promote intelligibility have been proposed, some inspired by human speech production changes, and useful gains in intelligibility in noise have been reported. The purpose of the current article is to evaluate within a common framework the performance of a

* Corresponding author at: Language and Speech Laboratory, Universidad del Pais Vasco, Vitoria, Spain. Tel.: +34 619616100.
E-mail address: m.cooke@ikerbasque.org (M. Cooke).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.01.001

range of speech modification strategies, alongside a number of natural speech styles.

Most speech modification algorithms proposed to date are noise-independent. Methods include boosting the consonant-vowel power ratio (Niederjohn and Grotelueschen, 1976; Skowronski and Harris, 2006; Yoo et al., 2007), spectral tilt flattening and formant enhancement (McLoughlin and Chance, 1997; Raitio et al., 2011), manipulation of duration and prosody (Huang et al., 2010), and voice conversion (Langner and Black, 2005). Raitio et al. (2011) also proposed a method using adaptation techniques (Yamagishi et al., 2009b) for hidden Markov model (HMM) text-to-speech (TTS) that require recordings of Lombard speech from the speaker whose voice is to be synthesized. A different approach, described by Moore and Nicolao (2011), also makes use of adaptation to map between normal, hypo, and hyper-articulated speech.

Some work has been carried out using prior knowledge or estimates of the noise context. These approaches include modification of the local signal-to-noise ratio (SNR) (Sauert and Vary, 2006; Tang and Cooke, 2010), optimisation of the spectral audio power reallocation based on the Speech Intelligibility Index (Sauert and Vary, 2010, 2011) or glimpse proportion (Tang and Cooke, 2012), cepstral extraction based on the glimpse proportion measure (Valentini-Botinhao et al., 2012a), and the insertion of small pauses (Tang and Cooke, 2011). Recently, Taal et al. (2012) presented an optimisation algorithm based on a spectro-temporal perceptual distortion measure.

The evaluation described in this paper aims to quantify the effect on intelligibility of speech modifications under energy and duration constraints. Listeners identified words in phonetically-balanced sets of utterances presented in both stationary and fluctuating maskers. Ten different types of speech were evaluated. These were either natural or synthetic speech, presented with and without modification. The two unmodified natural types – 'plain' and 'Lombard' – were produced in quiet and noise respectively. Five algorithmic modifications of natural speech were also evaluated, alongside unmodified synthetic speech and two further modified synthetic types.

The specific modification approaches selected for the evaluation were a subset of those developed in recent studies by the authors, chosen to exhibit a wide variety of potential modification techniques. Algorithms differed principally in their use of noise estimates, the parameters being modified, and the optimisation criterion employed. Alongside one noise-independent approach, others make use of information about the noise context during offline optimisation, while the rest employ online noise estimates. A number of the tested modification algorithms restrict themselves to changing spectral weights, either globally or locally in time; others additionally use time-domain amplitude range compression strategies. Some of the approaches were inspired by observed human speech production changes in intelligibility-enhancing types of speech, while others employed model-based optimisation of objective intelligibility.

The performance of each type is characterised in terms of the change in the percentage of keywords identified correctly by listeners. In addition, the concept of *equivalent intensity change* (EIC) is introduced, which describes the amount in decibels by which plain speech would need to be changed to acquire the same intelligibility as a given synthetic/modified type. A design goal for the evaluation was to be able to distinguish different speech types at a resolution of about 1 dB of EIC.

Section 2 provides a brief introduction to each of the 10 speech types evaluated in the current study. Speech and noise corpora are described in Section 3, along with details of the estimation of psychometric functions for the noise maskers. The outcome of the evaluation is presented in Section 4.

2. Speech types

Table 1 lists the 10 speech types whose intelligibility in noise is reported here, and summarises the extent to which each method uses noise signals or estimates both offline and online.

While the focus of the current study was on measuring intelligibility rather than naturalness or quality, informal listening suggested that most of the non-synthetic types were highly-natural and free from artefacts. The two methods which included a stage of dynamic range compression (SSDRC and TMDRC) were slightly less-natural than the others when screened in quiet, but when presented mixed with a masker these features were less noticeable.

2.1. Unmodified natural read speech (plain)

The reference unmodified speech type consisted of recordings of a subset of utterances from the Harvard sentence materials (Rothauser et al., 1969), which define 72 phonetically-balanced lists of 10 sentences each. Example sentences are "The key you designed will fit the lock" and "Open the crate but don't break the glass". These recordings provide a baseline to evaluate the intelligibility of unmodified speech. The type is referred to here as 'plain' speech, as suggested by Bradlow and Alexander (2007). However, it is worth noting that the speech was elicited via read sentences, so the plain type can be considered to consist of relatively clear speech. Further details of speech collection are given below in Section 3.1.

2.2. Speech spoken in the presence of a masker (Lombard)

Lombard speech (Lombard, 1911; Summers et al., 1988) refers to speech material elicited in the presence of noise. Lombard speech is of interest in evaluating speech intelligibility since it is a naturally modified speaking style which has been shown to be more intelligible than plain speech when presented at the same signal-to-noise ratio (Dreher

Table 1
Speech types tested.

Type	Approach	Mode	Modified	Noise dependency	
				Offline	Online
Plain	Neutral speech	Natural	No	–	–
Lombard	Lombard speech	Natural	No	–	–
OptSII	SII-optimised spectral reweighting	Natural	Yes	No	Short-term noise PSD
OptGP	Glimpse-optimised spectral reweighting	Natural	Yes	Yes	Noise type & SNR
SelBoost	Boost just audible regions	Natural	Yes	No	Short-term noise PSD
SSDRC	Spectral shaping + DRC	Natural	Yes	No	No
TMDRC	Harmonic model tilt modification + DRC	Natural	Yes	Yes	Noise type & SNR
TTS	HMM-based text-to-speech	Synthetic	No	–	–
TTSLomb	TTS adapted to Lombard	Synthetic	Yes	Yes	No
TTSGP	Glimpse-optimised TTS	Synthetic	Yes	No	Short-term noise PSD

and O'Neill, 1957; Summers et al., 1988; Pittman and Wiley, 2001; Lu and Cooke, 2008). Lombard sentences came from the same subset of the Harvard corpus as the plain material and were spoken by the same talker (see Section 3.1).

2.3. Speech Intelligibility Index-based optimisation of spectral audio power reallocation (OptSII)

In this approach, the audio power of the speech signal is spectrally reallocated with respect to the Speech Intelligibility Index (SII, ANSI S3.5-1997, 1997). A recursive closed-form optimisation scheme calculates, in each time frame, the spectral weights which maximise the SII given the current noise spectrum levels with the additional constraint of an unchanged short-term audio power of the speech signal. For this purpose, a (warped) filterbank with non-uniform frequency resolution divides speech and noise signal into 21 approximately Bark-scaled subband signals. This algorithm, which was originally designed for mobile communication devices, estimates all necessary information blindly based on the (far-end) speech and the (near-end) microphone signal with a look-ahead of only 10 ms. In situations where both participants speak at the same time (double-talk), the microphone signal contains not only background noise but also the speech signal of the local participant, which must not be treated as noise. This is addressed by a single-channel noise power spectral density (PSD) estimation algorithm. OptSII is thus noise-dependent, making use of online noise estimates. See Sauert and Vary (2011, 2010) for further details.

2.4. Optimal spectral reallocation based on glimpse proportion (OptGP)

As for OptSII technique, this approach focuses on promoting speech intelligibility in the context of different noise masker types and noise levels by applying a frequency-dependent weighting chosen to optimise the number of audible regions. The technique learns frequency band weights which maximise objective intelligibility using a genetic algorithm optimisation technique (Holland, 1975), with glimpse proportion (Cooke, 2006) as an objective intelligibility metric. Optimisation is performed separately for each masker type and SNR combination, resulting in a single weighting for each combination. A 58-channel auditory resolution filter bank is used rather than octave or third-octave band weights. Optimised weightings are typically almost binary in form, either boosting or attenuating individual frequency bands. Boosting becomes more sparse as noise level increases. This modification approach makes use of the masker signals during the offline learning process, but since a static spectral weighting is applied online, the method can be used based on a relatively high-level estimate of the noise context e.g., by estimation of the noise type and overall SNR. Further details are presented in Tang and Cooke (2012).

2.5. Selective energy reallocation to boost just audible time-frequency regions (SelBoost)

The SelBoost method is motivated by two factors: (i) some time-frequency regions are likely to possess a local SNR that is more than sufficient, while others are at or near the threshold of audibility; (ii) frequency regions differ in their importance for speech perception. For instance, the frequency range from 1000 to 4000 Hz is suggested to be more important than elsewhere (Zwicker, 1961; Studebaker et al., 1987; Bell et al., 1992). Under the constraint of constant input-output energy, transferring speech energy from regions of high local SNR to those of lower SNR, and from low-importance frequency regions to high-importance parts of the spectrum may be an effective strategy. The key offline requirement is to determine which parts of the local SNR range and which frequency bands are most effectively boosted by energy reallocation. An optimisation process which accomplishes this, reported in (Tang and Cooke, 2010), suggests boosting those regions whose local SNR is less than 5 dB in the frequency range 1800–7500 Hz. Moreover, this outcome was found not to depend on noise type or level. Online, this modification method requires good SNR estimates in each time-frequency region.

2.6. Spectral shaping and dynamic range compression (SSDRC)

SSDRC (Zorilă et al., 2012) performs spectral shaping followed by dynamic range compression (DRC). Spectral shaping consists of two cascaded subsystems which are adaptive to the probability of voicing: (i) an adaptive sharpening where the formant information is enhanced, and (ii) an adaptive pre-emphasis filter. Furthermore, a third fixed spectral shaping is used to prevent attenuation of high frequencies in the speech signal during signal reproduction. The operations of the spectral shaper follow observations of formant enhancement in clear speech (Hazan and Baker, 2011) and spectral tilt reductions in Lombard speech (Lu and Cooke, 2008). The output of the spectral shaping system is then input to the DRC, inspired by compression strategies used in sound recording and reproduction, audio broadcasting (Blesser, 1969) and amplification techniques in hearing aids (Kates et al., 1998). DRC has a dynamic and a static stage. During the dynamic stage, the envelope of the total time signal is dynamically compressed with a 2 ms release time constant and almost instantaneous attack time constant. During the static amplitude compression, the 0 dB reference level is set to 0.3 times the peak of the signal envelope. DRC enhances the transient components of speech. The SSDRC method is independent of noise type and level.

2.7. Harmonic model tilt modification and dynamic range compression (TMDRC)

TMDRC is a parametric approach based on full-band harmonic modelling of speech for increasing speech intelligibility in noise (Erro et al., 2012). This is achieved in two steps. First, the spectral slope is increased to mimic the effect of higher vocal effort. Then, the energy of the signal is redistributed over time to amplify low-energy parts of the signal. This is similar to the DRC operator discussed above. However, in this case the transformation operates on harmonic amplitudes and not on the envelope of the signal. It is interesting to note that the operations of spectral tilt modification and dynamic range compression can be easily integrated into a harmonic model-based waveform reconstruction module of a statistical text-to-speech synthesizer. This provides a higher degree of control on intelligibility at the expense of an almost negligible increment in computational load. TMDRC makes use of both offline and online noise estimates.

2.8. HMM-based text-to-speech synthesis (TTS)

Synthetic voices were built using the statistical and parametric HMM-based text-to-speech framework (Zen et al., 2009). The following parameters were used to train, adapt and generate speech: 59 Mel cepstral coefficients, Mel scale F0, and 25 aperiodicity energy bands extracted using STRAIGHT (Kawahara et al., 1999). A hidden semi-Markov model was used as the acoustic model. Observation vectors contained static, delta and delta-delta values, with one stream for the spectrum, three streams for F0 and one for the aperiodicity energy bands. The global variance method (Toda and Tokuda, 2007) was applied to compensate for the over-smoothing effect of acoustical modelling.

The standard synthetic speech material, referred to here as the speech type 'TTS', was created from a high quality average voice model (Yamagishi et al., 2009a) adapted to 2803 sentences from prior recordings made by the same male talker used for the current evaluation (see Section 3.1), corresponding to three hours of material. An average voice rather than a speaker-dependent voice was built because the plain speech dataset context coverage was not sufficiently large.

2.9. TTS adapted to Lombard speech (TTSLomb)

The Lombard voice 'TTSLomb' was based on TTS, further adapted using 780 sentences from the Lombard speech dataset described previously, corresponding to 53 mins of recorded material. Again, the reason for using adaptation was the lack of phonetic balance in the speech dataset. All acoustic features of the Lombard speech dataset i.e., Mel cepstral coefficients, F0, duration and band aperiodicity were used in the adaptation step for creating the TTSLomb voice. The Lombard voice produces sentences with longer duration (25% relative increase), longer pauses (18% relative increase), greatly increased F0 mean (39% relative increase) and flatter spectral tilt (24% flatter). TTSLomb uses offline noise estimates but is noise-independent at run time.

2.10. TTS optimised using a glimpse proportion metric (TTSGP)

To create the 'TTSGP' type a Mel cepstral coefficient modification method (Valentini-Botinhao et al., 2012b) was applied to the spectral parameters generated by the TTS type. Duration, fundamental frequency and excitation parameters remained unmodified. The first two Mel cepstral coefficients were modified (excluding the log-energy coefficient) in order to maximise intelligibility of speech in noise as given by an approximated version of the glimpse proportion measure (Cooke, 2006; Valentini-Botinhao et al., 2012a). The glimpse measure is maximised at each time frame, which means that there is no reallocation of energy across time frames, only within frequency regions. To calculate the measure and the Mel cepstral modification, a time-frequency auditory representation of speech and noise was extracted from the short term Fourier transform of the noise signal and the Mel cepstral coefficients of the speech signal. To extract this representation, 55 gammatone filters covering the range of 50–7500 Hz were used. Both convergence and distortion (10% relative increase in the Euclidian distance between the auditory representation of original and modified speech) were used as stopping

criteria. Like TTSLomb, the TTSGP type presents a flatter spectral tilt, though to a lesser degree (16%) than the TTSLomb type. TTSGP makes use of online noise estimates.

3. Methods

3.1. Speech material

A speech dataset comprised of natural sentences was chosen over the use of isolated words or restricted-vocabulary sentences in order to obtain evaluation results for phonetically-balanced materials more representative of everyday speech. The existing list of Harvard sentence materials (Rothauser et al., 1969) fits these criteria. The Harvard sentence lists define 72 sets of 10 sentences each. Each 10-sentence set is phonetically-balanced. Sets 1–18 (180 sentences) were used in the current evaluation.

A male native British English talker (Northern-influenced RP accent) working as a professional voice talent was recorded producing the entire set of 720 Harvard utterances in a hemi-anechoic chamber using two microphones: a Sennheiser HKH 800 p48 microphone on a stand, and a DPA 4035 headset microphone. Utterances were produced both in quiet and in the presence of a temporally-modulated speech-shaped noise masker (ICRA noise 5 from Dreschler et al., 2001), delivered over Beyerdynamic DT770 headphones at a calibrated level of 84 dB(A). In unmodified form, the utterances recording in quiet and noise are referred to as 'plain' and 'Lombard' as described in the previous section.

After recording, all speech was downsampled from 96 to 16 kHz using Praat (Boersma, 2001), manually endpointed to remove leading and trailing silence and high-pass filtered with a cut-off frequency of 100 Hz to remove low-frequency artefacts.

3.2. Maskers

In order to provide modification techniques with the opportunity to demonstrate sensitivity to both noise level and type, both fluctuating and steady-state maskers and a range of SNRs were tested. The fluctuating masker was competing speech (CS) from a female talker producing read news speech and Harvard-like sentences. The masker was generated by concatenating (in a random permutation fashion) 100 speech files produced by this talker. Concatenated speech was processed to eliminate all silences that were longer than 300 ms using SOX (SoX, 2012). The final processed file was 58 mins long. The steady-state masker was speech-shaped noise (SSN) whose long-term average speech spectrum matched that of the competing speech masker. SSN was generated by filtering white Gaussian noise through a 100th order all-pole filter which approximated the long-term spectrum of the CS masker. Note that the noise used to induce Lombard speech was not the same as either of the masker types but somewhat intermediate, sharing the temporal modulations of CS yet having a short-term spectrum of SSN. While different noise types induce acoustic modifications of varying degrees in Lombard speech (e.g., Lu and Cooke, 2008), the parameters affected by noise are largely independent of inducee noise type.

3.3. Speech-noise mixtures

Harvard sentences were centrally-embedded in CS and SSN masker fragments chosen at random from longer sequences. For SSN, a 30-s sample was generated, while for the CS masker extracts were drawn from the entire 58 mins waveform to reduce the probability of listeners being distracted by, or learning from, hearing the same background speech more than once. Rather than co-gating speech and noise (i.e., starting and ending both signals simultaneously), each masker fragment was one second longer than the sentence with which it was mixed, producing 0.5 s leading and lagging masker noise. Speech signals were padded with 0.5 s of inaudible low amplitude random noise at the beginning and end of each sentence. The reason for using non co-gated noise was to permit comparisons between modification approaches which produced speech of differing lengths, as explained below.

Speech-noise mixtures were created to measure the intelligibility of each of the 10 speech types[1]. Plain speech was added to noise at 3 SNRs, chosen to produce keyword scores of approximately 25, 50 and 75% (estimated in pilot tests to be −9, −4 and +1 dB for the stationary noise masker, and −21, −14 and −7 dB for the competing talker). In later sections these are referred to as 'Low SNR', 'Mid SNR', and 'High SNR'. To enable comparisons between plain and modified/generated speech types, signals were rescaled to produce the desired SNR. Since Lombard speech was on average longer than plain speech and to avoid constraining the intrinsic duration of synthetically-generated speech, modifications which resulted in durational increases within the additional 1s of masker were permitted. SNR calculation for modified speech, defined in Eqn. 1, is based on the modification-specific interval $[t_{1,m}, t_{2,m}]$ where the speech for type m is present:

$$10\log_{10}\frac{\sum_{t=t_{1,m}}^{t_{2,m}} s_m(t)^2}{\sum_{t=t_{1,m}}^{t_{2,m}} n(t)^2} = 10\log_{10}\frac{\sum_{t=t_1}^{t_2} s_{plain}(t)^2}{\sum_{t=t_1}^{t_2} n(t)^2} = \lambda \qquad (1)$$

where λ is the target SNR, $s_{plain}(t)$ and $s_m(t)$ are the plain and modified speech signals, and $[t_1, t_2]$ is the interval where the plain speech is present.

None of the types which resulted in modified durations (Lombard, TTS, TTSLomb, TTSGP) produced changes which exceeded the additional 1 s length of the masker signal.

[1] Two further modification algorithms were tested as part of the overall evaluation. However, these were based on prototype algorithms whose performance has subsequently substantially improved, so their results are not reported here.

3.4. Listeners

154 listeners were recruited using the University of Edinburgh's Student and Graduate Employment service. All were paid for their participation. All were young adults (age range predominantly 19–25) whose native language was English. All listeners received audiological screening, which led to responses from 15 listeners being removed from the analysis.

3.5. Design

Listeners identified keywords in speech in six conditions resulting from the combination of the two masker types presented at each of three SNR levels. 30 sentences were presented in each condition. Speech from each of 18 Harvard sets was mixed with noise for each of the six conditions, to produce 108 blocks of 30 sentences. Listeners were assigned to a subset of these blocks using a Latin square design which ensured that each listener heard one block in each of the six noise conditions, for a total of 180 sentences. No listener heard the same sentence twice, and each condition was heard by the same number of listeners. Within a single block the different speech types were mixed such that over the six blocks each listener heard the same number of sentences from each of the speech types. Prior to presentation, stimuli were normalised to have the same root mean square (RMS) level and 20 ms half-Hamming ramps were applies to attenuate onset and offset transients.

3.6. Procedure

Testing was performed in the School of Informatics at the University of Edinburgh, using individual sound-treated booths and Beyerdynamic DT770 headphones. Listeners were unable to modify output level. A custom-built MATLAB software application controlled the entire experiment. Listeners received two short practice sessions prior to the main test, one for each of the two masker types, presented at 0 dB SNR for SSN and −3 dB for CS, using Harvard sentences from outside the test subset. Each stimulus was presented once. Following presentation, listeners typed what they had heard, after which the subsequent stimulus was presented. In this way, the experiment was self-paced. Listeners needed 40–45 mins to complete the test.

Since few words were identifiable for some stimuli at the lower SNRs, listeners were instructed to simply type in the words they heard rather than to attempt to construct an entire sentence as their response. Null responses were not permitted by the software: listeners typed an 'X' for those sentences where no words were audible.

3.7. Scoring

Scores were computed based on the number of words correctly identified in each Harvard set. The short common words 'a', 'the', 'in', 'to', 'on', 'is', 'and', 'of', and 'for' were excluded. Prior to scoring, sentence lists and responses were adjusted to remove punctuation, and compound forms such as 'sideshow' or 'halfway' were modified to reflect the most common response type.

3.8. Estimation of psychometric functions

In order to be able to express the effect of modified speech types in terms of dB gains, a baseline listening test was designed to estimate the psychometric function which relates keyword scores to SNR for each of the two maskers. For the estimation of psychometric functions, plain speech was added to noise at 9 equally-spaced SNRs chosen to produce keyword scores in the range from 10 to 90%. For SSN the range was from −10.7 dB to 2.7 dB in 1.7 dB steps, while for CS the SNR varied from −28 dB to 0 dB in 3.5 dB steps.

Baseline testing was carried out prior to the main transcription test. 57 different listeners with a similar profile to that of the main test group (less eight listeners who did not pass the audiological test) identified keywords in speech in 18 conditions: two masker types crossed with each of nine SNR levels. Ten sentences were presented in each condition, each chosen as a single Harvard sentence block. Speech from each of 18 Harvard sets was mixed with noise for each of the 18 conditions, to produce 324 blocks of 10 sentences. Listeners were assigned to blocks of stimuli in a similar fashion to that used for the main test i.e., ensuring that no listener heard the same sentence more than once and that all listeners heard the same number of utterances in the 18 conditions. Testing took place in the same listening booths and under the same listening conditions as the main test. Scores were computed as for the main test.

Fig. 1 plots mean keyword scores in the baseline test as a function of SNR and masker type. The MATLAB `glmfit` function with a normal distribution and `logic` link function was used to find the best-fitting logistic function (Eq. 2) for each masker:

$$p_n = \frac{1}{1 + e^{-(\lambda - \mu_n)/s_n}} \tag{2}$$

where p_n is the proportion of words recognised correctly at SNR λ for masker n, and μ_n and s_n are the offset and slope of the logistic function for that masker.

Listeners tolerated around 9 dB more noise at the 50% words correct level for the competing speech masker compared to the speech-shaped noise masker, compatible with previous studies using stationary and fluctuating noise (Festen and Plomp, 1990).

4. Results

4.1. Keyword scores

Figs. 2 and 3 show keyword scores for the competing speech and speech-shaped noise maskers relative to scores

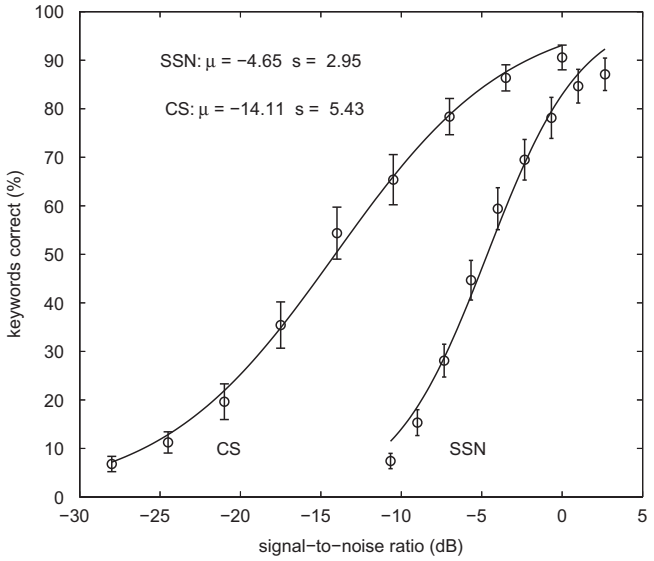

Fig. 1. Listeners' mean keyword scores (open circles) as a function of SNR for competing speech (CS) and speech-shaped noise (SSN) for 49 listeners. Error bars represent 95% confidence intervals. Two-parameter logistic fits are also shown.

for the plain speech type at each of three SNR levels denoted High, Mid, and Low. Speech types are ranked by degree of gain.

A 3-factor (modification, SNR level[2], masker type) repeated-measures ANOVA on arcsine-transformed keyword scores confirmed visual impressions, revealing highly-significant effects of all factors (all $p < 0.001$) as well as significant interactions between all factor combinations, suggesting that the effect of modification strategies varied across SNR and masker type.

Fisher's least significant differences, computed separately for each SNR level and masker type using ANOVAs with the single factor of modification type, are reported in Figs. 2 and 3 to allow statistical comparison of modification algorithms.

Leaving aside for the moment the three modifications which employed synthetic speech, most speech types were more intelligible than plain speech. The degree of gain showed significant variation across types, maskers, and SNR levels. On average, large gains were seen for modified speech in the presence of the stationary noise masker, reaching over 36 percentage points (from a plain speech baseline of around 16%). In general, the size of these gains was more than halved for the competing speech masker. Larger gains were observed at the more intense masker levels. In part, the scope for intelligibility improvements is limited in the High SNR condition, where for instance the near 8 percentage points gain of the most intelligible type was relative to a plain baseline of 86%. However, particularly in the CS case, room for further gains exists even at high SNRs.

The SSDRC method significantly outperformed all other approaches in four of the six conditions and was statistically-equivalent to the best method in the two High SNR conditions. Lombard speech was always more intelligible than plain speech, but it is notable that many of the manipulated types were even more intelligible.

TTS was always the least intelligible type (or statistically-equivalent to the type producing lowest scores), with a deficit relative to plain speech of up to 32 percentage points. However, the two modified synthetic types displayed substantial improvements relative to TTS, while not quite reaching the intelligibility of natural plain speech. In most conditions TTS adapted to Lombard speech (TTS-Lomb) outperformed TTS adapted to increase intelligibility (TTSGP).

4.2. Equivalent intensity change

While gains in keyword scores provide a raw measure of the effect of modifying speech, a more easily-interpretable measure of the effect of modifying speech is to estimate the amount by which plain speech would have to be boosted (or in some cases attenuated) to achieve the intelligibility level of the modified type. We compute a measure – the equivalent intensity change (EIC) – which represents the boost or attenuation level in decibels. The relation between EICs and changes in keyword scores is nonlinear. The EIC is derived by inverting the logistic approximation to the masker-specific psychometric function. The absolute SNR corresponding to the proportion of words correct in speech type m for masker n, denoted $p_{m,n}$, is computed using Eq. 3

$$\lambda_{m,n} = \mu_n - s_n \ln\left(\frac{1}{p_{m,n}} - 1\right) \tag{3}$$

where μ_n and s_n are the offset and slope of the psychometric function for masker n, whose specific values are provided in Fig. 1. The equivalent intensity change is then simply

$$EIC_{m,n} = \lambda_{plain,n} - \lambda_{m,n} \tag{4}$$

Note that the actual SNRs at which speech types were presented are not required in this calculation. Instead, via the logistic approximation, scores for all speech types (including plain speech) are derived from the same intelligibility-SNR relation. This approach implicitly corrects for any differences between the groups of listeners whose responses were used to estimate the psychometric functions and those who evaluated modified speech types.

Figs. 4 and 5 indicate that the most intelligible modifications produced around 5 dB of gain at Mid and Low SNRs for the stationary masker and 3–4 dB for the competing talker. Unmodified synthetic speech was 4–8 dB less intelligible than unmodified natural speech. However, some of this deficit was reduced for modified synthetic speech,

[2] Note that the factor levels for SNR – high, mid and low – do not represent the same SNRs (in dBs) for the two masker types.

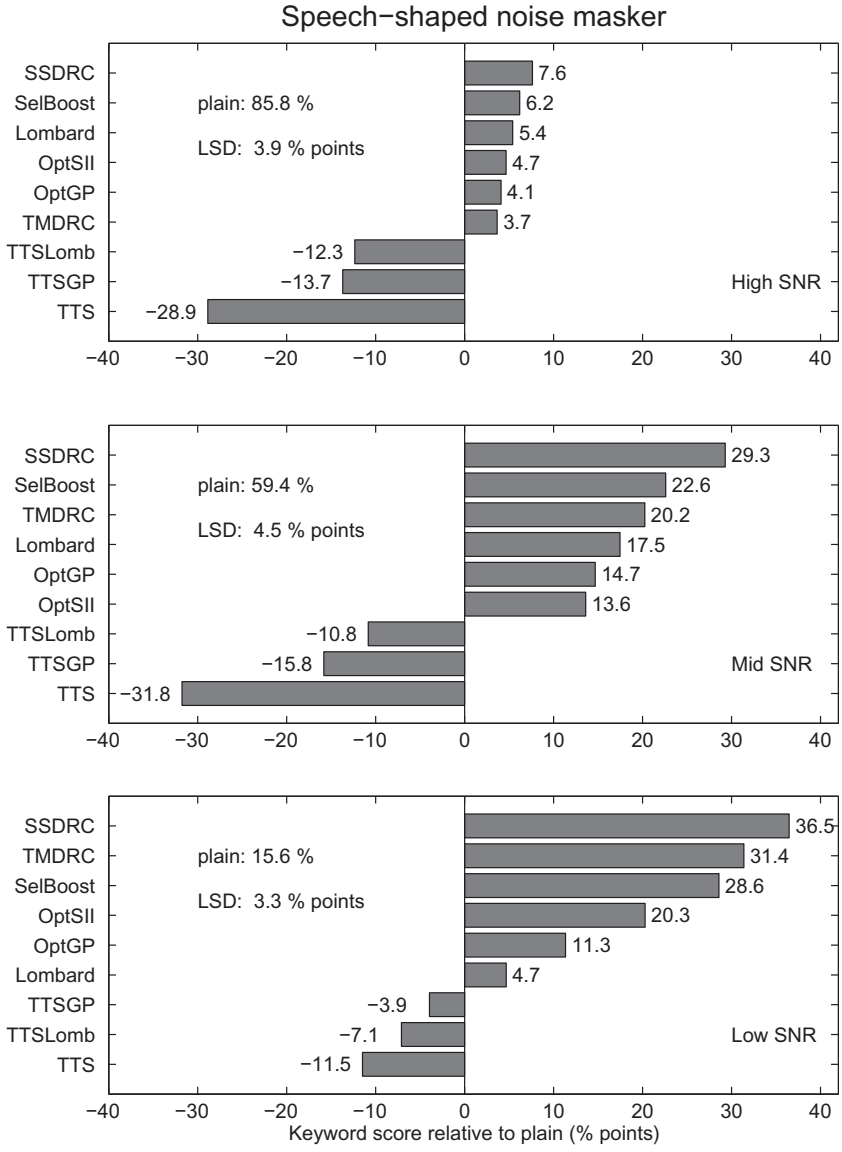

Fig. 2. Change in keywords correct scores for each modification type relative to natural plain speech, for the speech-shaped noise masker. Plain: absolute keyword score for plain. LSD: Fisher's least significant difference.

which produced gains of up to 4 dB compared to unmodified synthetic speech.

5. Discussion

5.1. Key findings

Speech modification can lead to substantial increases in intelligibility for sentences presented in noise relative to an unmodified speech baseline. The most successful techniques evaluated here produced increases in keyword scores over plain speech which ranged from 7.6 to 36.5 percentage points for a stationary masker, with smaller increases (5.5 to 15.4 percentage points) in the presence of a competing talker. These quantities correspond to gains in the range 2.5–5.2 dB and 2.4–4.1 dB for the two masker types. It is worth noting that these improvements are relative to a baseline of 'plain' speech, a read style produced by a professional voice talent. As such, on a continuum from conversational (hypoarticulated) to clear (hyperarticulated) speech, the plain type is already intrinsically quite clear and might therefore provide fewer opportunities for intelligibility gains. In this light, the outcome of the current evaluation of gains equivalent to boosting plain speech by up to 5 dB is encouraging.

In general, the size of the intelligibility gain increased with decreasing SNR. At the highest SNR, the scope for gains is limited by ceiling effects. In terms of relative reductions in word error rate, the largest change was observed at the moderate SNR for the speech-shaped noise masker, corresponding to a 72% relative reduction (i.e., an error rate of just over 40% for plain speech is reduced to around 11%).

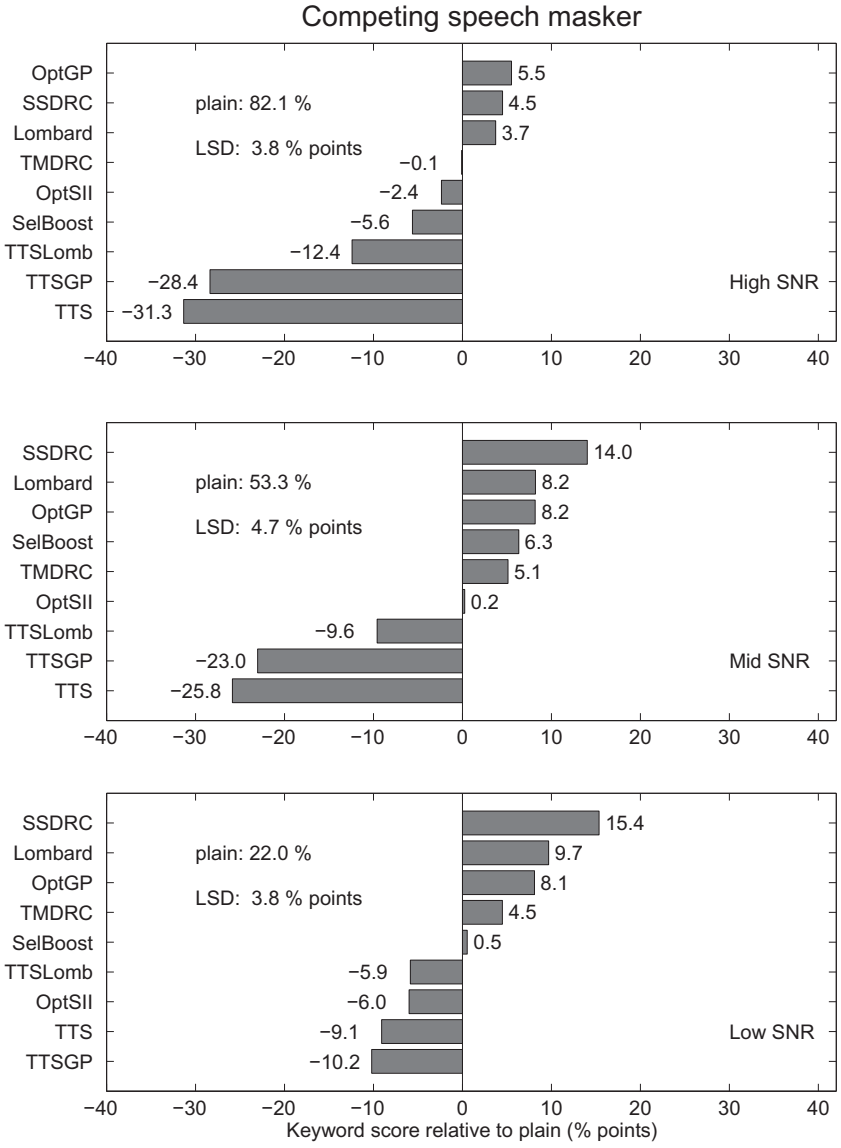

Fig. 3. As for Fig. 2 but for the competing speech masker.

Most modification techniques resulted in larger gains for the stationary masker than for the competing talker, suggesting that the algorithms evaluated here did not fully exploit modification opportunities afforded by temporal fluctuations in the masker. One technique – OptGP – did lead to larger gains for the competing talker at two of the three SNRs and was equivalent at the third SNR. We discuss possible reasons for this finding below.

In recent years, synthetic speech has reached intelligibility levels of natural speech in quiet (Yamagishi et al., 2008). However, the current study found a clear deficit in all noise conditions, ranging from 4 to 8 dB, suggesting that further work aimed at increasing the intelligibility of synthetic speech will be required to enable its adoption in a wider range of natural settings.

5.2. Comparison of algorithmic modifications to natural speech

Modified speech produced by SSDRC was significantly more intelligible than all other approaches in moderate and high levels of noise, for both maskers, and statistically equivalent to the best techniques in the low noise case. SSDRC was inspired by observations from talkers (clear and Lombard speech) and from audio engineering, and differs from most of the techniques evaluated here in modifying both spectral and temporal properties of the speech signal. The degree of spectral shaping is adaptive to the probability of voicing, with larger modifications for strongly-voiced speech segments. The adaptive character of spectral shaping is important to avoid introducing artifacts in the processed signal, especially in fricatives, silence, or other low-amplitude regions of speech. Concerning the

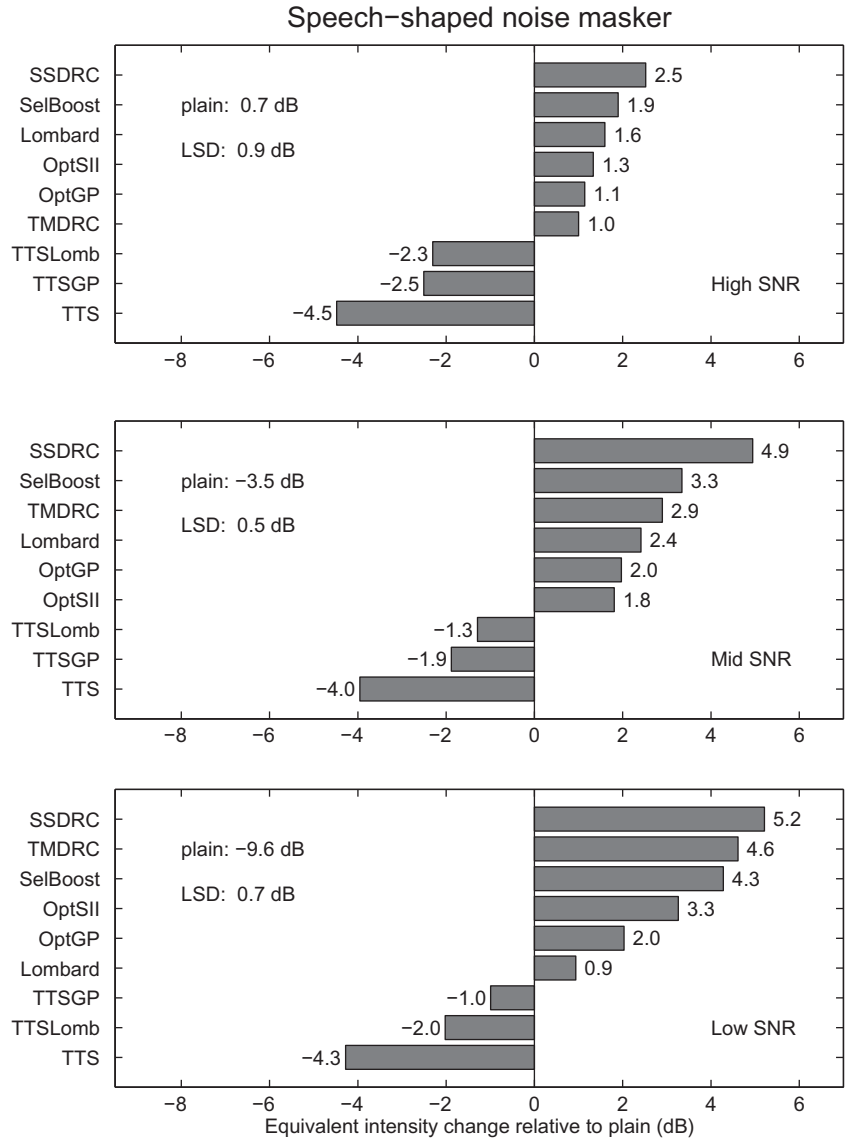

Fig. 4. Equivalent intensity change for each modification type relative to natural plain speech, for the speech-shaped noise masker. Plain: the dB value reported corresponds to the point on the psychometric function corresponding to the keyword score obtained for the plain type. LSD: Fisher's least significant different converted to dB via the psychometric function for this masker.

temporal element, dynamic range compression reduces the peak-to-RMS value of the speech signal, which has the effect of transferring energy from sonorant to less sonorant parts of speech. This type of compression supports observations (e.g., Hazan and Simpson, 1996; Yoo et al., 2007) which show that selective reinforcement of bursts, nasals, and vocalic onsets and offsets can provide significant improvements to the intelligibility of the subsequently degraded speech signal, while maintaining the same overall SNR. Enhancement of the transient components of speech has also been shown to improve intelligibility of speech in noisy conditions. Low energy signals are unaffected in order to avoid introducing artifacts in these areas of speech. Since the TMDRC method also employs a stage of dynamic range compression, the better performance of SSDRC can be attributed to differences in the treatment of spectral information. TMDRC is essentially a spectral tilt modifier, while SSDRC enhances formant information.

Two methods, OptSII and OptGP, used different objective intelligibility models to optimise speech modifications. These methods produced similar-sized moderate gains for two of three SNRs for the stationary masker. At the highest noise level, OptSII outperformed OptGP by 1.3 dB. In the competing talker case, OptGP produced larger gains than OptSII, with the difference ranging from 1.7 to 3.2 dB. The poorer performance of OptSII in the fluctuating masker case is easy to explain: the noise estimator is unable to track a rapidly-modulated masker such as competing speech. What is less clear is why the performance of OptGP is so good for this masker. Online, OptGP employs a static spectral weighting which is matched to the type of level of noise, based on offline optimisation.

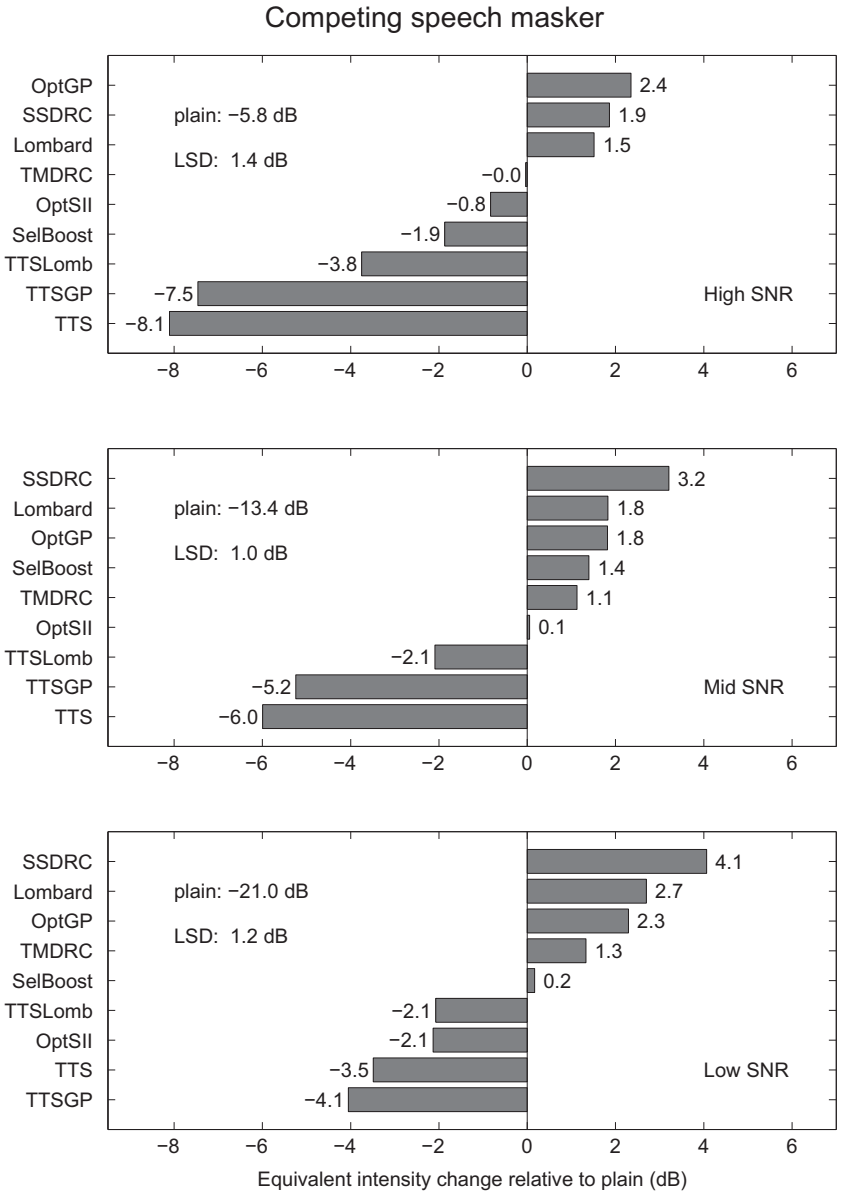

Fig. 5. As for Fig. 4 but for the competing speech masker.

Thus, OptGP cannot actively exploit the masker's temporal fluctuations. One possibility is that the sparse spectral weighting patterns learnt in OptGP (described in Tang and Cooke, 2012) create an energy concentration in narrow bands whose effect is to boost a small number of regions above the level of the masker for extended periods of time, enabling temporal variations in these frequency regions to be tracked by the listener. Having access to temporal modulations of the target speech is likely to be most beneficial when the masker itself is also modulated, since they help to define which frequency regions should be grouped together. Further work is needed to test this hypothesis.

The remaining natural speech modification approach, SelBoost, follows the pattern of benefits for the stationary masker and lack of substantial gains for competing speech. Like OptSII, SelBoost makes use of ongoing noise estimates to select which spectral regions to modify on a frame-by-frame basis. Unlike OptSII, SelBoost assumes access to ideal noise estimates, and might therefore be expected to show larger gains, particularly for the fluctuating masker. It is possible that boosting individual time-frequency cells leads to artefacts which have a negative influence on intelligibility. Since the competing speech masker is modulated in time and shows high variance across frequency, particularly in the resolved harmonics region, the probability of artefacts will be larger than for the stationary masker, compatible with the observed pattern of gains.

5.3. Lombard speech

In line with most studies (e.g. Dreher and O'Neill, 1957; Summers et al., 1988), Lombard speech was always more

intelligible than plain speech, although the difference was marginal for some conditions. Lombard is a naturally modified style, and observations from Lombard speech formed part of the motivation for several techniques, either indirectly (SSDRC, TMDRC) or directly (TTSLomb), so it is interesting that many of the artificially-modified approaches led to larger increases in intelligibility than Lombard speech. However, the Lombard benefit was more equally-balanced overall across the two masker types than was the case for many algorithmic modifications. Since Lombard speech was produced in response to temporally-modulated speech-shaped noise, it is notable that gains are observed in the presence of mismatched maskers, i.e., competing speech and stationary speech-shaped noise.

5.4. Synthetic speech

While synthetic speech (TTS) was substantially less intelligible than plain natural speech, the other TTS-based approaches led to very significant gains over the TTS baseline, ranging from 2.2 to 3.3 dB for the SSN masker and 2.0 to 4.3 dB for the competing speaker. In most conditions, these gains took synthetic speech to within 1–2 dB of natural speech in noise. In fact, the benefit of adapting TTS to Lombard speech was larger than the difference between natural Lombard and plain speech in five of the six conditions. The two TTS modifications are motivated by quite different criteria—adapting to Lombard and optimising an objective intelligibility model respectively. While the first modification changes spectral features, duration, and prosody, the latter only modifies the speech spectral envelope, hinting at a possible beneficial combination of the two approaches.

5.5. Practical considerations: noise estimation and computational complexity

As summarised in Table 1, modification methods differed in the extent to which they made use of information from the masker. Two approaches (TMDRC, OptGP) used noise samples in offline training and assumed knowledge of either or both of noise type and SNR online. TTSLomb had a similar offline noise dependency, but no dependency online. Other methods such as OptSII, SelBoost and TTSGP required ongoing noise estimates at the frame level. Intriguingly, the only fully noise-independent approach, SSDRC, was the one that produced the largest gains, suggesting that even larger EICs might be produced by extensions of SSDRC which make use of the noise-context.

In terms of computational cost, SSDRC, TMDRC, OptGP and OptSII are of low complexity and can be performed in real time. SSDRC is a frame-based non-parametric approach involving simple estimation and filtering operations, while OptGP is a stationary spectral weighting. The TTSLomb approach has the same complexity as the baseline TTS approach, while the TTSGP approach requires more FFT operations for the gradient calculation performed in each time frame.

5.6. Extensions

We conclude by outlining possible extensions to both the modification algorithms and the evaluation methodology which may inform future evaluations.

- Some elements of the algorithmic modifications tested here can be combined. For instance, dynamic range compression can be applied as a post-modification process at the level of the modified speech signal, for both natural and synthetic speech.
- Optimisation using other objective intelligibility or quality models (e.g. Christiansen et al., 2010; Taal et al., 2011; Rix et al., 2001) is likely to result in different modifications.
- To accommodate natural durational differences between plain, Lombard, and synthetic speech, constrained changes in overall duration were permitted (see Section 3.3), although none of the modified natural types took advantage of this. Likewise, no technique tested within-utterance redistribution of segment durations. This is a clear area for future work, given the role of segment duration in cueing phonological distinctions such as voicing. Similarly, apart from Lombard adaptation in TTSLomb, none of the techniques explored modifications to parameters such as F0.
- Baseline natural styles other than plain and Lombard speech could be used, including speech produced by explicit instructions to speak clearly, casual speech, and Lombard speech induced by more complex maskers such as competing talkers.
- Criteria other than constant input-output SNR such as equal-loudness constraints may be desirable.
- Metrics other than intelligibility include those typically employed in the evaluation of processed and synthetic speech e.g., quality, naturalness, and comprehensibility. In addition, there is scope for measures which explore the effect of modified speech at a cognitive level (e.g., multitasking, cross-modal effects).
- The effect of speech modifications for different listener groups (e.g., non-native listeners, listeners with hearing impairment) could be considered.
- Similarly, the size of any benefits of modification is likely to vary across talkers.
- One limitation of the current evaluation is the use of headphone presentation of speech and noise. In many applications listeners will have access to other cues such as those which carry spatial information and which can help to separate speech from noise. The additional benefit of modified speech needs to be evaluated in conditions found in each scenario (e.g., using loudspeaker presentation in real environments). Although a recent

study by Raitio et al. (2012) in different test environments concluded that realistic noise environments are not a prerequisite for intelligibility testing of synthetic speech, they did find that the use of a variety of setups yielded additional information about the effect of noise on speech.

6. Conclusions

This paper reports the results of the first large-scale evaluation of speech production modification strategies designed to increase intelligibility in noise without changing overall signal-to-noise ratio. Some modification approaches were inspired by studies of human speech modes known to be intelligible, while others sought modifications which optimised one of several objective intelligibility models. A number of modification algorithms led to useful gains, equivalent to increasing the level of unmodified but intrinsically rather clear speech by up to 5 dB. Gains were observed for natural and synthetic speech in both stationary and fluctuating maskers. These outcomes establish that speech modification is a highly-effective strategy for promoting message reception in adverse conditions.

Acknowledgements

We thank Vasilis Karaiskos for help in running the listening tests, Julian Villegas for contributions to the recording of speech material, and T-C. Zorilă, V. Kandia and D. Erro for useful discussions on developing SSDRC and TMDRC. 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 by the Future and Emerging Technologies (FET) programme under FET-Open grant number 256230 (LISTA).

References

ANSI S3.5-1997, 1997. Methods for the calculation of the Speech Intelligibility Index.

Bell, S.T., Dirks, D.D., Trine, T.D., 1992. Frequency-important functions for words in high- and low context sentences. J. Speech Hear. Res. 35, 950–959.

Blesser, B.A., 1969. Audio dynamic range compression for minimum perceived distortion. IEEE Trans. Audio Electroacoust. 17.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Internat. 5, 341–435.

Bradlow, A.R., Alexander, J.A., 2007. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. J. Acoust. Soc. Amer. 121, 2339–2349.

Christiansen, C., Pedersen, M.S., Dau, T., 2010. Prediction of speech intelligibility based on an auditory preprocessing model. Speech Comm. 52, 678–692.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Amer. 119, 1562–1573.

Cooke, M., Lu, Y., 2010. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Amer. 128, 2059–2069.

Dreher, J., O'Neill, J., 1957. Effects of ambient noise on speaker intelligibility for words and phrases. J. Acoust. Soc. Amer. 29, 1320–1323.

Dreschler, W.A., Verschuure, H., Ludvigsen, C., Westermann, S., 2001. ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment. Audiology 40, 148–157.

Erro, D., Stylianou, Y., Navas, E., Hernaez, I., 2012. Implementation of simple spectral techniques to enhance the intelligibility of speech using a harmonic model. In: Proc. Interspeech, Portland, USA .

Festen, J., Plomp, R., 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J. Acoust. Soc. Amer. 88, 1725–1736.

Hazan, V., Baker, R., 2011. Acoustic–phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. J. Acoust. Soc. Amer. 130, 2139–2152.

Hazan, V., Simpson, A., 1996. Cue-enhancement strategies for natural VCV and sentence materials presented in noise. Speech Hear. Lang. 9, 43–55.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

Howell, P., Barry, W., Vinson, D., 2006. Strength of British English accents in altered listening conditions. Percept. Psychophys. 68, 139–153.

Huang, D.Y., Rahardja, S., Ong, E.P., 2010. Lombard effect mimicking. In: Proc. SSW7, Kyoto, Japan, pp. 258–263.

Kates, J.M., 1998. Signal processing for hearing aids. In: Kahrs, M., Brandenberg, K. (Eds.), Applications of Digital Signal Processing to Audio and Acoustics. Kluwer Academic Publishers, Boston, pp. 235–277.

Kawahara, H., Masuda-Katsuse, I., Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Comm. 27, 187–207.

Langner, B., Black, A.W., 2005. Improving the understandability of speech synthesis by modeling speech in noise. In: Proc. ICASSP, pp. 265–268.

Lindblom, B., 1990. Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, pp. 403–439.

Lombard, E., 1911. Le signe d'élévation de la voix (the sign of the elevation of the voice). Annales des maladies de l'oreille et du larynx 37, 101–119.

Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble and stationary noise. J. Acoust. Soc. Amer. 124, 3261–3275.

McLoughlin, I.V., Chance, R.J., 1997. LSP-based speech modification for intelligibility enhancement. In: Proc. Digital Signal Processing, Santorini, Greece, pp. 591–594.

Moore, R.K., Nicolao, M., 2011. Reactive speech synthesis: Actively managing phonetic contrast along an H&H continuum. In: 17th Internat. Cong. on Phonetic Sciences, pp. 1422–1425.

Niederjohn, R.J., Grotelueschen, J.H., 1976. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. IEEE Trans. Acoust. Speech Signal Process. 24, 277–282.

Patel, R., Schell, K.W., 2008. The influence of linguistic content on the Lombard effect. J. Speech Lang. Hear. Res. 51, 209–220.

Picheny, M., Durlach, N., Braida, L., 1985. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. J. Speech Hear. Res. 28, 96–103.

Pittman, A.L., Wiley, T.L., 2001. Recognition of speech produced in noise. J. Speech Lang. Hear. Res. 44, 487–496.

Raitio, T., Suni, A., Vainio, M., Alku, P., 2011. Analysis of HMM-based Lombard speech synthesis. In: Proc. Interspeech, Florence, Italy, pp. 2781–2784.

Raitio, T., Takanen, M., Santala, O., Sun, A., Vainio, M., Alku, P., 2012. On measuring the intelligibility of synthetic speech in noise do we

need a realistic noise environment? In: Proc. ICASSP, pp. 4015–4028.

Rix, A., Beerends, J., Hollier, M., Hekstra, A., 2001. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. In: Proc. ICASSP, pp. 749–752.

Rothauser, E.H., Chapman, W.D., Guttman, N., Silbiger, H.R., Hecker, M.H.L., Urbanek, G.E., Nordby, K.S., Weinstock, M., 1969. IEEE Recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17, 225–246.

Sauert, B., Vary, P., 2006. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In: Proc. ICASSP, Toulouse, France, pp. 493–496.

Sauert, B., Vary, P., 2010. Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement. In: Proc. ITG-Fachtagung Sprachkommunikation, Bochum, Germany.

Sauert, B., Vary, P., 2011. Near end listening enhancement considering thermal limit of mobile phone loudspeakers. In: Proc. Conf. on Elektronische Sprachsignalverarbeitung (ESSV), Aachen, Germany, pp. 333–340.

Skowronski, M.D., Harris, J.G., 2006. Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. Speech Comm. 48, 549–558.

SoX, 2012. SoX–Sound eXchange. Software. Available [Apr. 2012] from <http://sox.sourceforge.net/>.

Studebaker, G.A., Pavlovic, C.V., Sherbecoe, R.L., 1987. A frequency important function for continuous discourse. J. Acoust. Soc. Amer. 81, 1130–1138.

Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., Stokes, M., 1988. Effects of noise on speech production: Acoustic and perceptual analysis. J. Acoust. Soc. Amer. 84, 917–928.

Taal, C.H., Hendriks, R.C., Heusdens, R., 2012. A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. In: Proc. ICASSP, pp. 4061–4064.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19, 2125–2136.

Tang, Y., Cooke, M., 2010. Energy reallocation strategies for speech enhancement in known noise conditions. In: Proc. Interspeech, pp. 1636–1639.

Tang, Y., Cooke, M., 2011. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In: Proc. Interspeech, Florence, Italy, pp. 345–348.

Tang, Y., Cooke, M., 2012. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In: Proc. Interspeech, Portland, USA.

Toda, T., Tokuda, K., 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. Inf. Syst. E90-D, 816–824.

Uther, M., Knoll, M.A., Burnham, D., 2007. Do you speak E-NG-L-I-SH? a comparison of foreigner- and infant-directed speech. Speech Comm. 49, 2–7.

Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., Zen, H., 2012. Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise. In: Proc. ICASSP Kyoto, Japan, pp. 3997–4000.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2012. 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.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009a. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. 17, 66–83.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.H., Toda, T., Tokuda, K., King, S., Renals, S., 2009b. A robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Trans. Audio Speech Lang. Process. 17, 1208–1230.

Yamagishi, J., Zen, H., Wu, Y.J., Toda, T., Tokuda, K., 2008. Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In: Proc. Blizzard Challenge Workshop.

Yoo, S.D., Boston, J.R., El-Jaroudi, A., Li, C.C., Durrant, J.D., Kovacyk, K., Shaiman, S., 2007. Speech signal modification to increase intelligibility in noisy environments. J. Acoust. Soc. Amer. 122, 1138–1149.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Comm. 51, 1039–1064.

Zorila, T.C., Kandia, V., Stylianou, Y., 2012. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In: Proc. Interspeech, Portland, USA.

Zwicker, E., 1961. Subdivison of audible frequency range into critical bands. J. Acoust. Soc. Amer. 33, 248.