INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The effects of modified speech styles on intelligibility for non-native listeners

Martin Cooke[1,2]*, Maria Luisa Garcia Lecumberri*[2]

[1]Ikerbasque (Basque Science Foundation), Bilbao, Spain
[2]University of the Basque Country, Vitoria, Spain
m.cooke@ikerbasque.org, garcia.lecumberri@ehu.es

Abstract

Speech output, including modified and synthetic speech, is used increasingly in natural settings where message reception might be affected by noise. Recent evaluations have demonstrated the effect of different speech styles on intelligibility for native listeners, but their impact on listening in a second language is less well-understood. The current study measured the intelligibility of four speech styles in the presence of stationary and fluctuating maskers for a non-native listener cohort, and compared the results with those of native listeners on the same task. Both groups showed a similar pattern of effects, but the scale of intelligibility gains and losses with respect to plain speech was significantly compressed for the non-native group relative to native listeners. In addition, non-native listeners identified speech from the four styles in the absence of noise, revealing that styles shown to be beneficial in noise lost their benefits or were harmful in quiet conditions. This result suggests that while enhanced styles lead to gains by reducing the effect of masking noise, the same styles distort the acoustic-phonetic integrity of the speech signal. More work is needed to develop speech modification approaches that simultaneously preserve speech information and promote unmasking.

Index Terms: Modified speech, synthetic speech, Lombard effect, speech perception, non-native listeners

1. Introduction

With an increasing use of speech output technology, listeners more frequently encounter styles of speech which differ from what might be considered the norm, namely plain natural speech. Non-canonical styles include synthetic speech and recorded natural speech which has been modified with the goal of enhancing intelligibility e.g. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. These forms of speech are sometimes deployed in adverse acoustic environments such as public transport interchanges or over bandlimited telephone channels. Recent large-scale evaluations with participants listening in their native language have demonstrated significant benefits of modified natural speech in adverse conditions [11]. Since listening in a non-native language can be considered as a further form of adverse condition [12], it is of interest to know how synthetic and modified forms of speech interact with noise for non-native listeners (NNL). The current study examines the impact of non-canonical styles of speech on intelligibility for this listener group relative to native listeners (NL).

Collectively, previous studies have produced an inconsistent picture of the effect of differing speech styles on intelligibility in noise for NNLs. In [13], for synthetic speech in noise, NNLs suffered proportionally more than a native cohort. Speech produced in noise, so-called Lombard speech [14], did not benefit NNLs to quite the same extent as NLs [15]. Similarly, NNLs benefitted significantly less than NLs from a clear speech style in noise [16].

On the other hand, [17] found similar degrees of benefit for spectral filtering and selective speech enhancements for one native and two non-native listener groups whose task was to identify consonants in nonsense vowel-consonant-vowel material in stationary noise. More recently, NNLs exposed to 6 different kinds of modified speech exhibited almost identical changes over an unmodified baseline as NLs using simple sentences [18] in stationary and modulated maskers. For clear speech, [19] demonstrated similar benefits for NLs and a cohort of high-proficiency NNLs.

Comparing these disparate studies is, of course, complicated by the range of second languages and competences of the NNL groups, and the use of different speech material, maskers and tasks. It is known that such factors have an important bearing on the differential performance of NL and NNLs in adverse conditions (see review in [20]). One goal of the current study is to measure the relative intelligibility of differing speech styles using a common non-native listener cohort with the same sentence material in the presence of identical maskers.

The use of non-native listeners provides an opportunity to explore an additional question. The studies cited above are based on speech in noise, typically used to avoid ceiling effects for native listeners. Consequently, it is difficult to distinguish the source of any intelligibility increases produced by a given speech style: do listeners benefit because a given speech style is intrinsically clearer (e.g., less reduced), or is the speech style simply more resistant to masking? Non-native listener groups provide a way to tease apart these two explanations, since, unlike native listeners, they typically perform at well below ceiling levels in quiet conditions. The value of this approach was demonstrated in a recent study of the effect of Lombard speech on NNLs [15], which showed that Lombard speech was less intelligible than non-Lombard speech in quiet conditions, in spite of the fact that the same speech material was significantly more intelligible in noise. A second goal of the current study is to address the 'intrinsic' vs. 'maskability' confound by measuring the effect of different speech styles on non-native listeners in quiet as well as in masked conditions.

Non-native listeners identified keywords in everyday sentences presented in quiet and in two types of masking noise, one stationary, the other fluctuating, at two signal-to-noise ratios (SNRs). Four styles of speech were compared: unmodified natural speech, Lombard speech, speech algorithmically-modified to enhance intelligibility, and synthetic speech. Results in the masked conditions were compared to those of native listeners who identified identical stimuli in a recent evaluation of modified speech styles [11].

http://dx.doi.org/10.21437/Interspeech.2016-41

2. Methods

2.1. Speech materials

Speech material consisted of the first 180 sentences of the Harvard Corpus [21] produced in the following four styles:

- **Plain** An unmodified speech style consisting of sentences read by a male British English speaker. Although the Plain style represents a baseline for speech modification approaches, it is highly-intelligible and can be considered as a form of clear speech.
- **Lombard** To generate Lombard speech, the same talker was also recorded producing the same sentences in the presence of a temporally-modulated speech-shaped noise (ICRA noise 5, [22]), delivered over headphones at a level of 84 dBA.
- **SSDRC** This condition consists of Plain sentences processed by the Spectral Shaping and Dynamic Range Compression technique [23]. SSDRC involves a sequence of processes which enhance formant frequencies, apply adaptive pre-emphasis, and supply a fixed boost to high frequencies. The SSDRC approach produced the largest gains in the evaluation described in [24], with intelligibility increases equivalent to a reduction in SNR of up to 5.2 dB.
- **TTS** TTS sentences were generated by an HMM-based text-to-speech synthesis framework [25] which made use of an average voice model [26] adapted to 2803 sentences (3 hours of speech) from the same male talker who spoke the Harvard sentences.

Figure 1 shows spectrograms of a sentence produced in the four styles. Further details can be found in sections 2.1, 2.2, 2.6 and 2.8 respectively of the native listener study [11].

2.2. Maskers

A steady-state speech-shaped noise (SSN) and a temporally-fluctuating competing speech (CS) masker, both employed in [27], were used in the current study. The competing talker was a female speaker reading newspaper speech and Harvard-like sentences. The SSN masker had a long-term spectrum matching that of the competing talker.

In the native listener study [11] participants were tested at three SNRs in each masked condition. Since pilot studies suggested that the most adverse SNR for each masker resulted in very low scores for NNLs, only the high and middle SNRs, referred to as 'snrHi' and 'snrMid', were used. These SNRs were chosen to produce approximately 75% and 50% keywords correct amongst native listeners. In addition, non-native listeners heard the sentences in a noise-free condition (Table 1).

In the Plain condition, speech-plus-noise mixtures were produced by centrally-embedding the target sentence in a masker fragment with 500 ms of lead and lag. That is, the total duration of the masker exceeded that of the target speech by 1 second. Since durations were potentially different in the remaining three speech styles, these conditions resulted in different amounts of masker lead and lag. Token-wise SNR was computed over the region where the speech was present.

Table 1: *Maskers and SNRs.*

masker		CS		SSN	
condition	quiet	snrHi	snrMid	snrHi	snrMid
SNR	∞	-7 dB	-14 dB	+1 dB	-4 dB

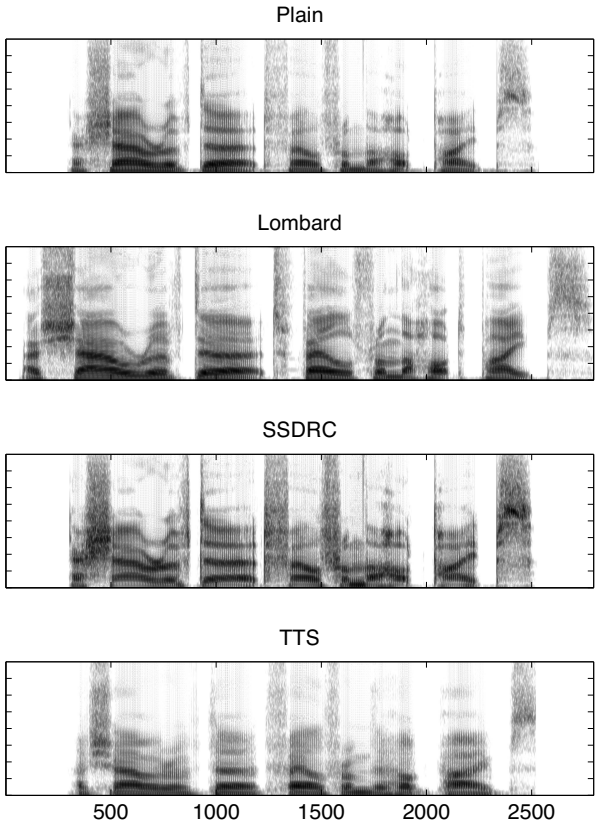

Figure 1: *Spectrograms of "A shower of dirt fell from the hot pipes" for each speech style. Note that Lombard speech typically has longer segments than the other styles. The effect of dynamic range compression can be seen in the SSDRC spectrogram in boosting the weak fricatives /f/ at around 1500 and 1750 ms.*

2.3. Listeners

A cohort of 44 normal-hearing listeners participated in the experiment. All were native monolinguals in Spanish or bilingual in Spanish and Basque, and all were in their second year of studies for the degree of English Philology at the University of the Basque Country. Listeners were paid for their participation.

2.4. Procedure

Each listener heard all 180 sentences, 60 in each of the quiet, snrMid and snrLo conditions. For the non-quiet conditions, listeners heard equal numbers of sentences in the presence of the CS and SSN maskers. Sentences were blocked by masker and SNR, leading to 6 blocks of 30 sentences (i.e., two blocks corresponding to the quiet condition). Within each block, listeners heard either 7 or 8 examples of each of the four speech styles. Assignation of speech styles to blocks was carried out in such a way as to ensure that no listener heard the same sentence more than once, and each sentence was heard the same number of times in each speech style and masker condition. Block ordering was balanced across listeners.

Listeners were tested individually in a sound-attenuating acoustic booth in the Phonetics Laboratory at the University of the Basque Country. Stimuli were presented at a fixed comfortable listening level over Sennheiser HD 650 headphones. The experiment was self-paced under computer control, listening typing their responses into an on-screen text entry box. Prior to the main experiment, listeners heard 6 practice utterances in each of the two masker conditions.

2.5. Postprocessing

All words apart from 'a','the','in','to','on','is','and','of' and 'for' were scored as keywords. No manual spelling corrections were carried out. Scores are presented as the percentage of keywords correct in each condition. For statistical analysis, percentages were converted to rationalised arcsin units [28].

3. Results

3.1. Keyword scores in noise

Figure 2 shows keyword identification rates for NNLs as a function of speech style, masker and SNR. The two modified styles expected to show intelligibility gains, Lombard and SSDRC, do indeed result in increased scores relative to the Plain baseline. In the SSN masker, gains of between 7-9 and 17-18 percentage points (p.p.) were produced for Lombard and SSDRC respectively. More modest increases of between 2-5 and 5-7 p.p. can be seen for the CS masker. Synthetic speech was always less intelligible than Plain speech, with drops of 6-9 p.p. A similar pattern across styles is evident at both SNR levels.

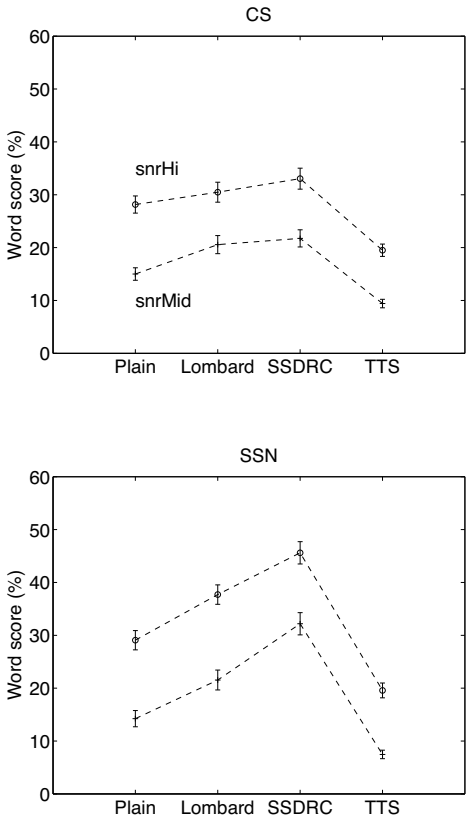

Figure 2: *Keyword scores across speech styles/SNRs. Top: competing speaker masker; bottom: speech-shaped noise.*

Separate repeated-measures ANOVAs with factors of speech style and SNR support visual perceptions, with clear effects of both SNR [CS: $F(1,43) = 288, p < 0.001, \eta^2 = 0.24$; SSN: $F(1,43) = 258, p < 0.001, \eta^2 = 0.3$] and speech style [CS: $F(3,129) = 44.9, p < 0.001, \eta^2 = 0.2$; SSN: $F(3,129) = 127, p < 0.001, \eta^2 = 0.42$] and no interaction between the two factors for either CS [$p = 0.4$] or SSN [$p = 0.33$].

For SSN, the intelligibility of all speech styles were significantly different from each other at both SNRs [Fisher's Least Significant Difference of 4.0 and 4.6 RAU for snrHi and sndMid respectively]. However, in the CS case, the Lombard and SSDRC styles were equally intelligible at both SNRs [Fisher's LSD: 3.8 and 4.3 RAU for snrHi and sndMid]. In addition, in the snrHi condition, Lombard speech was equivalent in intelligibility to the Plain style.

3.2. Keyword scores in quiet

Figure 3 presents keyword scores in the absence of masking noise. Non-native listeners scored well below ceiling in this task. Intelligibility clearly differs across speech style [$F(3,129) = 18.3, p < 0.001, \eta^2 = 0.07$]; apart from Lombard and Plain, all styles differ [Fisher's LSD: 2.8 RAU]. These results suggest that when noise is not present, any putative gains from the two modified speech styles that were beneficial in noise (Lombard, SSDRC) disappear. Indeed, in the case of SSDRC there is a significant loss in intelligibility.

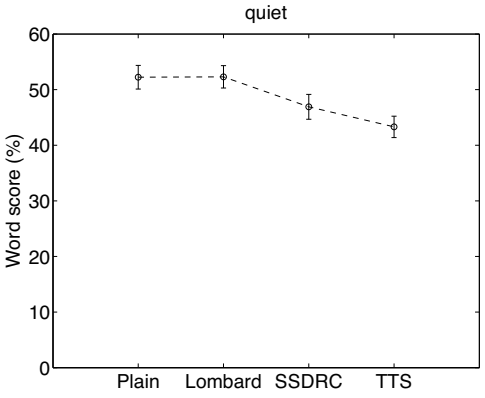

Figure 3: *Keyword scores in quiet.*

3.3. Comparison with native listener scores

Figure 4 compares native and non-native listener keyword scores and gains over the Plain style for shared conditions (snrHi and snrMid); native data is from [11]. The two sets of data are clearly well-correlated: for both scores and gains, $\rho = 0.92$ [$p < 0.001$]. However, non-native listeners performed well below native listeners on this task. For NNLs, average gains were around 44% of those produced by NLs. By the same token, in the TTS conditions the losses suffered by NNLs were less than those of NLs. Note that native listeners in the snrHi condition were already scoring at a high level, so the relatively small gains observed in the lower panel of Fig. 4 probably represent a ceiling effect.

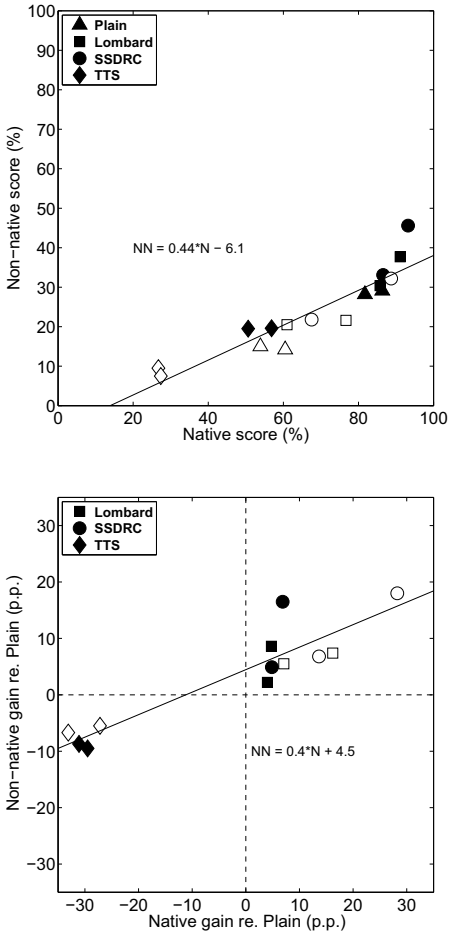

Figure 4: *Native and non-native scores (top) and gains over Plain (bottom). Filled: snrHi; open: snrMid; maskers are not distinguished. Solid lines shows the best linear fits.*

4. Discussion

Here, different speech styles induced a very similar pattern of intelligibility for native and non-native listeners when identifying keywords in sentences masked by noise. We speculate that this finding mainly reflects the ability of each speech style to withstand energetic masking, since the amount of information available following the interaction of speech and masker signals is the same for NLs and NNLs. This notion is supported by the contrasting pattern of performance across speech styles in the absence of noise. SSDRC-modified speech, while highly-beneficial in noise, leads to a drop in intelligibility in quiet, suggesting that gains in noise do not come from improvements to the acoustic-phonetic saliency of speech, with energetic masking release as the most likely cause of the masked benefit. Indeed, the loss in quiet might result from damage to the acoustic-phonetic structure of speech following SSDRC processing.

While the two listener groups show a similar pattern, NNLs identified substantially fewer keywords in each condition, generally showing a *multiplicative* intelligibiity loss relative to NLs, represented by the linear model NNL $= 0.44 * \text{NL} - 6.1$. This finding contrasts markedly with our earlier study [18] where NNL scores were well-predicted by a near *additive* relationship NNL $= 0.97 * \text{NL} - 12.5$ i.e. a small but constant loss relative to natives. The biggest difference between the two studies is in the nature of the target sentence material. In [18] listeners identified the letter-number combination (e.g. "G 2") from matrix sentences of the form "place red at G 2 now". This type of sentence places low demands on vocabulary and syntax, has a highly-predictable pattern, and, while not entirely absent, contains fewer of the rich contextual cues that are known to benefit native listeners in adverse conditions [20].

The reduced effectiveness of both Lombard and SSDRC styles in the presence of the competing speech masker has been observed in other evaluations [11, 24]. Obtaining intelligibility gains in a temporally-fluctuating masker such as CS may depend on exploiting epochs of favourable local SNR. Since the Lombard speech was induced by a different fluctuating masker, it is likely that the gains produced in the CS condition were less than optimal. SSDRC makes no use of information about the masker, so it is not surprising that gains were lower for CS.

Synthetic speech, in line with other studies e.g. [29, 30], was less intelligible than natural speech (although *modified* TTS can exceed natural speech intelligibility [31]). An examination of the spectrograms in Figure 1 reveals that TTS suffers from a less well-defined formant structure, particularly during transients (e.g., the diphthong in "shower"), and there are clear segment durational differences between TTS and Plain styles that may have contributed to reduced intelligibility.

Lombard speech was moderately-beneficial to NNLs, consistent with the findings of [15]. It is worth noting that the level of the noise used to induce Lombard speech in the current study was relatively low. In the Lombard speech induced by a similar mild noise condition of [15], NNLs showed equivalent performance to speech produced in quiet, when tested in noise-free conditions, echoing the findings of the current study. However, in [15], Lombard speech induced by a more intense masker, which was more effective than mild Lombard speech in noise, was less intelligible in quiet. That outcome parallels the finding with SSDRC in the current study, supporting the notion that speech styles designed to overcome masking noise can be harmful when heard outside the noise context.

More generally, the measured intelligibility of a given speech style presumably reflects the net effect of both advantageous and disadvantageous changes with respect to a Plain style. This raises the possibility that if modified speech styles were to be developed with a focus on preserving acoustic-phonetic integrity, the potential gains in masked conditions could exceed those seen to date.

5. Conclusions

Native and non-native listeners showed a similar pattern of intelligibility change when confronted by differing speech styles in noise. However, NNLs suffered a multiplicative change with respect to NLs, unlike earlier studies with simpler sentences where an additive change has been observed. Speech styles beneficial in noise led to losses, or no gains, in quiet conditions, suggesting that while certain styles promote energetic masking release, they do so at the expense of acoustic-phonetic integrity.

Acknowledgements: We thank Yannis Stylianou for providing code implementing the SSDRC algorithm, and Cassia Valentini-Botinhao and Junichi Yamagishi for providing the TTS sentences. This work was supported in by the Basque Government Consolidado grant to LASLAB and the Spanish Ministry MICIN Diacex project.

6. References

[1] I. V. McLoughlin and R. J. Chance, "LSP-based speech modification for intelligibility enhancement," in *Proc. Digital Signal Processing*, vol. 2, Santorini, Greece, Jul. 1997, pp. 591–594.

[2] 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.

[3] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 493–496.

[4] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Comm.*, vol. 48, no. 5, pp. 549–558, 2006.

[5] D. Bonardo and E. Zovato, "Speech synthesis enhancement in noisy environments," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2853–2856.

[6] S. D. Yoo, J. R. Boston, A. El-Jaroudi, C.-C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1138–1149, Aug. 2007.

[7] D. Erro, Y. Stylianou, E. Navas, and I. Hernaez, "Implementation of simple spectral techniques to enhance the intelligibility of speech using a harmonic model," in *Proc. Interspeech*, Portland, USA, 2012, pp. 639–642.

[8] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, Portland, USA, September 2012, pp. 955–958.

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

[10] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *Proc. ICASSP*, Mar. 2012, pp. 4061–4064.

[11] 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. 55, pp. 572–585, 2013.

[12] S. Mattys, M. H. Davis, A. R. Bradlow, and S. K. Scott, "Speech recognition in adverse conditions: A review," *Lang. Cognitive Proc.*, vol. 27, no. 7-8, pp. 953–978, Sep. 2012.

[13] M. Reynolds, Z. Bond, and D. Fucci, "Synthetic speech intelligibility: comparison of native and non-native speakers of English," *Augmentative and Alternative Communication*, vol. 12, pp. 32–36, 1996.

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

[15] M. Cooke and M. García Lecumberri, "The intelligibility of Lombard speech for non-native listeners," *J. Acoust. Soc. Am.*, vol. 132, pp. 1120–1129, 2012.

[16] A. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *J. Acoust. Soc. Am.*, vol. 112, pp. 272–284, 2002.

[17] V. Hazan and A. Simpson, "The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects," *Language and Speech*, vol. 43, pp. 273–294, 2000.

[18] M. Cooke, M. García Lecumberri, and Y. Tang, "The effect of speech modification on non-native listeners for matrix-style sentences," *J. Acoust. Soc. Am.*, vol. 137, pp. EL151–EL157, 2015.

[19] R. Smiljanic and A. Bradlow, "Bidirectional clear speech perception benefit for native and high-proficiency non-native talkers and listeners: Intelligibility and accentedness," *J. Acoust. Soc. Am.*, vol. 130, pp. 4020–4032, 2011.

[20] M. García Lecumberri, M. Cooke, and A. Cutler, "Non-native speech perception in adverse conditions: A review," *Speech Comm.*, vol. 52, no. 11-12, pp. 864–886, 2010.

[21] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

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

[23] T. C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, Portland, USA, 2012, pp. 635–638.

[24] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[25] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[26] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, pp. 66–83, 2009.

[27] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech and Language*, vol. 28, pp. 543–571, 2014.

[28] G. Studebaker, "A rationalized arcsine transform," *J. Speech Hear. Res.*, vol. 28, pp. 455–462, 1985.

[29] H. Venkatagiri, "Segmental intelligibility of four currently used text-to-speech synthesis methods," *J. Acoust. Soc. Am.*, vol. 113, pp. 2095–2104, 2003.

[30] E. Axmear, J. Reichle, M. Alamsaputra, K. Kohnert, K. Drager, and K. Sellnow, "Synthesized speech intelligibility in sentences: a comparison of monolingual English-speaking and bilingual children," *Language, Speech and Hearing Services in Schools*, vol. 35, pp. 244–250, 2005.

[31] C. Valentini-Botinhao, J. Yamagishi, S. King, and Y. Stylianou, " Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise," in *Proc. Interspeech*, 2013, pp. 3567–3571.