Objective measures for predicting the intelligibility of spectrally smoothed speech with artificial excitation

Danny Websdale, Thomas Le Cornu and Ben Milner

University of East Anglia

{d.websdale, t.le-cornu, b.milner}@uea.ac.uk

10.21437/Interspeech.2015-219

Abstract

A study is presented on how well objective measures of speech quality and intelligibility can predict the subjective intelligibility of speech that has undergone spectral envelope smoothing and simplification of its excitation. Speech modifications are made by resynthesising speech that has been spectrally smoothed. Objective measures are applied to the modified speech and include measures of speech quality, signal-to-noise ratio and intelligibility, as well as proposing the normalised frequency-weighted spectral distortion (NFD) measure. The measures are compared to subjective intelligibility scores where it is found that several have high correlation ($|r| \geq 0.7$), with NFD achieving the highest correlation ($r = -0.81$).

Index Terms: intelligibility, spectral smoothing, STRAIGHT

1. Introduction

The aim of this work is to examine how well objective measures of speech quality and speech intelligibility can predict the intelligibility of speech that has undergone modification to its spectral envelope and excitation. Spectral envelope modifications take the form of smoothing where spectral detail is gradually reduced. Excitation modifications consider effects such as a monotone pitch, an artificial time-varying pitch and a wholly unvoiced excitation.

Our motivation comes primarily from the area of audio-visual speech processing where several studies have considered estimating audio speech features from visual speech [1, 2]. In such situations it is difficult to estimate spectrally detailed audio features and instead the estimated spectral envelope is highly smoothed. It is therefore useful to know how spectral smoothing affects the intelligibility, and how effective objective measures are at predicting the reduction in intelligibility. Furthermore, visual speech provides no voicing or fundamental frequency information and so it is interesting to examine the impact on intelligibility with no knowledge of the speech excitation. The study has wider application, for example in the area of hearing loss, where some types of hearing impairment are attributed to a broadening of auditory filters. This can be likened to spectral smoothing which makes perception of spectral shape difficult for listeners leading to reduced intelligibility [3, 4].

The approach taken is similar to a study that examined how well objective measures predict the intelligibility of noisy speech [5]. Using human listeners, speech intelligibility was measured under different noise conditions and speech enhancement methods and its correlation with various objective measures analysed. The objective measures included methods for predicting speech quality (e.g. PESQ [6]), signal-to-noise ratio (SNR) measures and intelligibility measures, and found large variations in their correlation to intelligibility. PESQ and coherence-based measures [7] had high correlation, while LPC-based measures and SNR had lower correlation. Several similar studies have examined the ability of objective measures to predict speech quality and intelligibility under varying noise conditions and speech distortions for normal hearing and hearing impaired listeners [8, 9, 10].

The remainder of the paper is organised as follows. The speech database, application of speech modifications and the organisation of the subjective tests are described in Section 2. A brief description of the objective measures is provided in Section 3 as well as an introduction of a proposed measure, the normalised frequency-weighted distortion measure (NFD). Section 4 analyses the correlation between the objective and subjective measures under the various speech modifications.

2. Method

This section describes the preparation of the speech data and the set up for the subjective and objective intelligibility testing.

2.1. Speech database

The GRID database was used for the intelligibility tests and contains recordings from 34 speakers who each produced 1000 sentences [11]. Each sentence comprises six words and follows the grammar shown in Table 1. One male and one female speaker were used for the evaluations and the audio was downsampled to 8 kHz.

Table 1: GRID sentence grammar.

command	colour	preposition	letter	digit	adverb
bin	blue	at	A-Z	1-9	again
lay	green	by	minus W	zero	now
place	red	in			please
set	white	with			soon

2.2. Speech modification

Two kinds of distortion are applied to the speech – a smoothing of the spectral envelope and a change in the excitation in terms of its voicing and fundamental frequency. These manipulations were applied by decomposing the speech into its source and filter components, then applying the respective modifications before reconstructing a speech signal. The STRAIGHT vocoder was chosen for reconstruction based on its success in hidden Markov model-based speech synthesis [12, 13]. STRAIGHT requires three parameters: fundamental frequency, f_{0_i}, a time-frequency surface, $X(f, i)$, and aperiodicity, $A(f, i)$, where i and f are the frame number and frequency bin.

2.2.1. Spectral envelope smoothing

Two methods for smoothing the spectral envelope have been applied: LPC and filterbank. For both methods the speech is first segmented into 50 % overlapping 20 ms frames. Pth order LPC analysis is applied to each frame and the coefficients transformed into a smooth spectral envelope to create a smoothed spectral surface, $X^{\mathrm{LPC}}(f,i)$, for STRAIGHT. Filterbank smoothing uses a K-channel mel filterbank applied to the magnitude spectrum of each frame. This is interpolated to the dimensionality of the time-frequency surface to give $X^{\mathrm{FB}}(f,i)$.

2.2.2. Voicing and fundamental frequency

Four variants of voicing and fundamental frequency are considered and summarised in Table 2. For aperiodicity, setting $\tilde{A}(f,i)$ to $-\infty$ produces entirely periodic speech, while setting to 0 produces entirely aperiodic speech. *Original* uses

Table 2: f_0 and aperiodicity modifications.

Method	$\tilde{\mathrm{A}}(f,i)$	$\tilde{f}_{0_i}$
Original	$-\infty$	f_{0_i}
Monotone	$-\infty$	μ_{f_0}
Time-varying	$-\infty$	$\mu_{f_0} + \Delta_{f_0}\, cos((2\pi i \,/\, 400) + \phi_r)$
Unvoiced	0	–

the original voicing and fundamental frequency. The three artificial methods assume no knowledge of the voicing of the speech. *Monotone* uses a constant fundamental frequency of either 103 Hz or 216 Hz, which is the mean f_0 for the male and female speaker respectively. *Time-varying* modulates the mean fundamental frequency by a sinusoid function with a time period of 4 s, and a range about the mean, Δ_{f_0}, of ± 17.5 Hz or ± 28 Hz for the male and female speaker. A random phase offset, ϕ_r, is also added to vary the starting f_0 value. *Unvoiced* applies unvoiced excitation to the entire utterance.

The modified time-frequency surface, $X^{\mathrm{LPC}}(f,i)$ or $X^{\mathrm{FB}}(f,i)$, aperiodicity, $\tilde{A}(f,i)$, and fundamental frequency, $\tilde{f}_{0_i}$, are input into STRAIGHT to create each test utterance.

2.3. Listening test description

Five levels of LPC smoothing, P = {2, 4, 6, 8, 14}, and five sizes of filterbank, K = {4, 7, 10, 15, 20}, were evaluated. These gave a range of smoothing from almost no distortion to highly distorted. Each smoothing variant was combined with each of the four excitation variants to give 40 configurations. The original unprocessed speech was also included in the listening tests. These were repeated for both speakers to give a total of 82 configurations.

Twenty listeners took part in the subjective tests and each heard speech from the 82 configurations. The listening tests were conducted in a quiet environment using headphones and listeners recorded the words they heard. For each of the 82 configurations, the percentage of words correctly recognised by the listeners was averaged and forms the intelligibility value.

3. Objective measures

A range of objective measures of speech quality and speech intelligibility are considered, as well as measures for hearing impaired listeners. A new measure, the NFD, designed specifically to measure the intelligibility of smoothed speech is introduced.

3.1. PESQ and LPC-based measures

PESQ is a commonly-used objective measure that has high correlation to subjective quality ($r > 0.9$) across a range of test conditions [14, 9]. It also correlates well to the intelligibility of noisy and enhanced speech ($r = 0.79$) [5].

Two LPC-based measured are considered: log likelihood ratio (LLR) and cepstral distance (CEP) [15, 16]. These measure differences in spectral envelope between the original and modified speech,

$$\mathrm{LLR} = \frac{1}{N}\sum_{i=0}^{N-1} \log\left(\frac{\boldsymbol{a}_i^m \boldsymbol{R}_i^o (\boldsymbol{a}_i^m)^T}{\boldsymbol{a}_i^o \boldsymbol{R}_i^o (\boldsymbol{a}_i^o)^T}\right) \tag{1}$$

$$\mathrm{CEP} = \frac{1}{N}\sum_{i=0}^{N-1} \frac{10}{\log(10)} \sqrt{2\sum_k [c_i^o(k) - c_i^m(k)]^2} \tag{2}$$

$\boldsymbol{a}_i^o$, $\boldsymbol{R}_i^o$ and $\boldsymbol{c}_i^o$ are the LPC vector, autocorrelation matrix and cepstral vector of the original speech, while $\boldsymbol{a}_i^m$ and $\boldsymbol{c}_i^m$ are computed from the modified speech. N is the number of frames in the utterance.

3.2. Signal-to-noise ratio

Of the various measures of SNR, the frequency-weighted segmental SNR (fwSNRseg) [9] is considered based on its high correlation to intelligibility reported in [5], and computed,

$$\mathrm{fwSNRseg} = \frac{10}{N}\sum_{i=0}^{N-1} \frac{\sum_{j=1}^{K} W(j,i) \log_{10} \frac{(X^o(j,i))^2}{(X^o(j,i) - X^m(j,i))^2}}{\sum_{j=1}^{K} W(j,i)} \tag{3}$$

$X(j,i)^o$ and $X(j,i)^m$ are critical band magnitude spectra in the jth band of the original and modified signals and $W(j,i)$ is a band importance function, which is discussed in Section 3.6.

3.3. Intelligibility measures

This study considers four objective intelligibility measures. The normalised covariance metric (NCM), the short-term articulation index (AI-ST) and the coherence speech intelligibility index (CSII) [17, 5, 7] are computed,

$$\mathrm{NCM} = \frac{\sum_{j=1}^{K} W(j) \times TI^{\mathrm{NCM}}(j)}{\sum_{j=1}^{K} W(j)} \tag{4}$$

$$\mathrm{AI{-}ST} = \frac{1}{N}\sum_{i=0}^{N-1} \frac{\sum_{j=1}^{K} W(j,i) \times TI^{\text{AI-ST}}(j,i)}{\sum_{j=1}^{K} W(j,i)} \tag{5}$$

$$\mathrm{CSII} = \frac{1}{N}\sum_{i=0}^{N-1} \frac{\sum_{j=1}^{K} W(j,i) \times TI^{\mathrm{CSII}}(j,i)}{\sum_{j=1}^{K} W(j,i)} \tag{6}$$

These are frequency weighted summations of transmission index (TI) values computed within frequency bands. The methods differ in how the TIs are computed and in the band importance functions, W. TIs within each frequency band are computed from normalised SNR measures, with TI^{NCM} computed from the cross-covariance of bandpass filtered signals from the original and modified signals. $TI^{\text{AI-ST}}$ is computed from the SNR of the modified signal while TI^{CSII} is computed from a signal-to-distortion ratio using the magnitude squared coherence [18].

The fourth measure is the short-time objective intelligibility measure (STOI) [19] which begins by extracting a time-frequency surface based on a 15 channel filterbank from the

original and modified utterances, $X_i^o(j)$ and $X_i^m(j)$. The intelligibility of a single time-frequency (TF) unit, $d_i(j)$, is computed from the L preceding TF units by computing the correlation between the TF units from original and modified signals,

$$d_i(j) = \frac{\sum_l \left(X_l^o(j) - \mu_{X_j^o}\right)\left(X_l^m(j) - \mu_{X_j^m}\right)}{\sqrt{\sum_l \left(X_l^o(j) - \mu_{X_j^o}\right)^2 \sum_l \left(X_l^m(j) - \mu_{X_j^m}\right)^2}} \quad (7)$$

The means, $\mu_{X_j^o}$ and $\mu_{X_j^m}$, are computed from the M preceding TF units which, together with the denominator, serves to normalise the correlation measurement across the range of TF units used to compute $d_i(j)$. Finally, STOI is computed by averaging $d_i(j)$ across all frequency bands and frames

$$\text{STOI} = \frac{1}{N \times 15} \sum_{i=0}^{N-1} \sum_{j=1}^{15} d_i(j) \quad (8)$$

3.4. HASQI and HASPI

Objective measures have been developed to predict the quality and intelligibility of degraded speech for listeners with hearing loss and using hearing aids. Speech distortions encountered in these situations include dynamic range compression, frequency shifting, envelope modifications, spectral smoothing and time-varying gain, and so are interesting candidates for analysis. The two measures considered are the Hearing Aid Speech Quality Index (HASQI) and Hearing Aid Speech Perception Index (HASPI) [10, 20, 21]. HASQI comprises two measures, a nonlinear term, Q_{nonlin}, that is affected by noise and nonlinear distortion, and a linear term, Q_{lin}, that is affected by linear filtering and spectral changes. HASPI combines measures of temporal fine structure and time-frequency envelope from the original and modified signals to give a measure of intelligibility.

3.5. Normalised frequency-weighted distortion measure

We propose an additional objective measure that is designed to measure spectral envelope distortion. The original and modified signals are first segmented into 50 % overlapping 20 ms duration frames and their spectral envelopes, $\grave{X}_i^o(j)$ and $\grave{X}_i^m(j)$, obtained by computing their respective cepstra, lowpass liftering and then returning to the power spectrum [22]. A band importance function, $W(j)$, is then applied to the spectral envelopes to give frequency-weighted spectral envelopes, $\tilde{X}_i^o(j)$ and $\tilde{X}_i^m(j)$

$$\tilde{X}_i^o(j) = W(j)\grave{X}_i^o(j) \quad \text{and} \quad \tilde{X}_i^m(j) = W(j)\grave{X}_i^m(j) \quad (9)$$

Rather than computing the mean square error across the entire utterance between $\tilde{X}_i^o(j)$ and $\tilde{X}_i^o(j)$, normalisation is first applied by subtracting, for each frame, the mean of the frequency weighted spectral envelope, μ_i^o and μ_i^m, which are computed across frequency for each frame. This gives the normalised frequency-weighted spectral distortion (NFD) measure

$$\text{NFD} = \frac{1}{NK} \sum_{i=0}^{N-1} \sum_{j=1}^{K} \left[\left(\tilde{X}_i^o(j) - \mu_i^o\right) - \left(\tilde{X}_i^m(j) - \mu_i^m\right)\right]^2 \quad (10)$$

3.6. Band importance function

The band importance function, W, can be predefined (e.g. AI weights [23]) or take on signal-dependent values. Studies reported superior results when using signal-dependent weights as this allows values to change dynamically as the signal and distortion change [5]. Functions that assign more weight in bands with higher signal energy were found to give higher correlation. A range of BIFs are investigated in the analysis in Section 4.2.

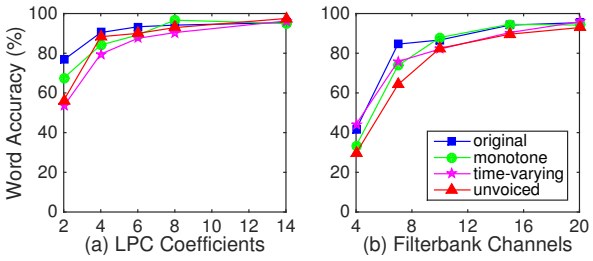

Figure 1: *Mean word accuracy for (a) LPC smoothing, (b) Filterbank smoothing, using original, monotone or time-varying fundamental frequency and unvoiced excitation.*

4. Experimental results

Experiments first examine the effect of spectral envelope and excitation modification on subjective intelligibility. Secondly, a correlation analysis of the subjective and objective intelligibility is presented.

4.1. Subjective measures

Figure 1(a) shows subjective intelligibility (word accuracy) obtained when smoothing the spectral envelope using LPC analysis with order P from 2 to 14 and the effect of using the original, monotone, time-varying and unvoiced excitation. Similarly, Figure 1(b) shows intelligibility using filterbank smoothing with 4 to 20 channels.

Intelligibility remains largely constant from 14 to 8 LPC coefficients and then falls with fewer coefficients. With only 2 coefficients, which gives just one spectral peak, intelligibility is 75 % using the original excitation and 70 % with an artificial (monotone) excitation. This is surprising considering other studies have reported a general requirement of two spectral peaks for vowel classification [24], however, this difference is attributed to the intelligibility tests in this paper being whole word-based within a constrained grammar. Filterbank smoothing is more damaging to intelligibility. This is attributed to filterbank channel frequencies not necessarily being positioned at spectral peaks, whereas with the LPC analysis the envelope fits the frequency and bandwidth of spectral peaks to provide a better spectral representation. Moving from 20 to 7 channels reduces intelligibility by 10 % which falls rapidly to 30-40 % with 4 channels. At this point the peaks in the spectral envelope are significantly different to those in the original spectral envelope.

The tests show variation in intelligibility using the different excitation methods. As may be expected, the original excitation gives highest intelligibility. When less smoothing is applied, artificial excitation has little effect on intelligibility. At higher levels of smoothing, artificial excitation tend to reduce intelligibility, in some cases by as much as 20 %.

4.2. Objective measures

Pearson's correlation coefficient, r, is used to measure the correlation between the objective and subjective intelligibility scores averaged across the 82 configurations. The standard deviation of the error, $\sigma_e = \sigma_d\sqrt{1-r^2}$, where σ_d is the standard

Table 3: Correlation coefficient and standard deviation of the error between word accuracy and objective measures.

Measure	Band importance function	r	σ_e
PESQ	–	0.63	0.15
fwSNRseg	$W(j,i) = X(j,i)^p, p = 8$	0.61	0.15
LLR	–	-0.63	0.15
CEP	–	-0.65	0.14
NCM	$W(j) = 1$	0.70	0.13
AI-ST	$W(j,i) = X(j,i)^p, p = 11$	0.44	0.17
CSII	$W(j,i) = 1$	0.22	0.18
$\text{CSII}_{\text{high}}$	$W(j,i) = 1$	0.24	0.18
CSII_{mid}	$W(j,i) = 1$	0.30	0.18
CSII_{low}	$W(j,i) = X(j,i)^p, p = 1$	0.44	0.17
STOI	–	0.75	0.12
$\text{HASQI}_{\text{nonlin}}$	–	0.62	0.15
$\text{HASQI}_{\text{lin}}$	–	0.30	0.18
$\text{HASQI}_{\text{comb}}$	–	0.58	0.15
HASPI	–	0.64	0.14
NFD	SII [23, Table B.1]	-0.81	0.11

deviation of the subjective intelligibility scores is also calculated. Table 3 shows r and σ_e for the objective measures in Section 3 and the best band importance function where used.

The measures traditionally used to measure speech quality (PESQ, fwSNRseg, LLR and CEP) perform almost equally well with correlations in the range $|r| = 0.61$ to 0.65. Of the intelligibility measures, even with an exhaustive search of BIFs, the various CSII variants and AI-ST perform poorly ($r \leq 0.44$) while NCM and STOI are substantially higher ($r \geq 0.7$). Of the measures developed for hearing impaired listeners, HASPI and the nonlinear component of HASQI perform almost equally ($r = 0.62, 0.64$), while the linear component is much less effective ($r = 0.30$). Highest performance is found with NFD ($r = -0.81$). Considering all measures with $|r| \geq 0.7$, these have similarities in that they normalise the signals over which differences are computed, in terms of either their means and/or standard deviations, and so effectively concentrate on fluctuations in spectral envelope which appears to correlate well with intelligibility.

Considering the two best performing objective measures, STOI and NFD, Figure 2 shows scatter plots for male and female speech, with the symbols showing the method of excitation as indicated from they key in Figure 1. Both measures maintain high correlation for both genders with NFD having higher correlation than STOI, which is less effective on the female speaker. This is attributed to a wider range in STOI values at high levels of subjective intelligibility, compared to NFD which has lower variation. Considering the different methods of excitation, high correlation is maintained with little variation found across the different methods when measuring excitation specific correlation values.

Figure 3 shows scatter plots for STOI and NFD separated into LPC and filterbank smoothing methods. NFD maintains high correlation across both methods of smoothing while STOI reduces for LPC smoothing. This is again attributed to STOI values having wider variation at high subjective intelligibility than observed for NFD. High correlation can also be observed when considering the individual methods of speech excitation.

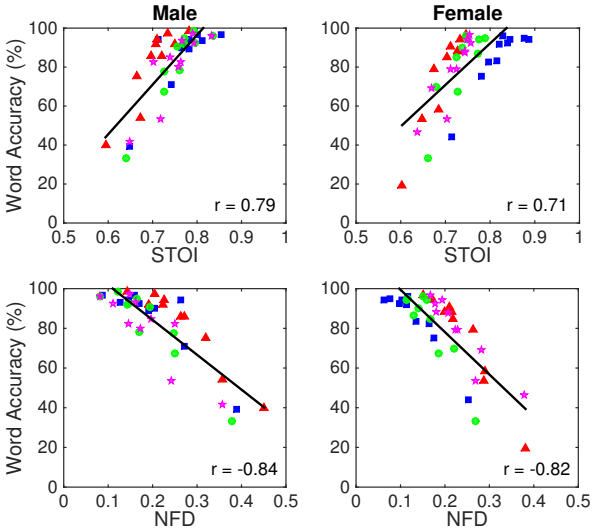

Figure 2: *Scatter plots comparing subjective intelligibility with STOI and NFD, for male and female speakers.*

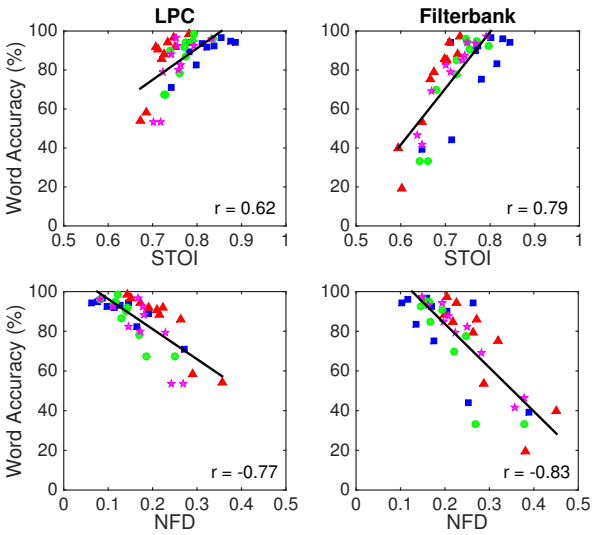

Figure 3: *Scatter plots comparing subjective intelligibility with STOI and NFD, for LPC and filterbank smoothing.*

5. Conclusions

Subjective listening tests revealed that speech that has been highly smoothed, even down to just one spectral peak, remains largely intelligible. Replacing the original excitation with highly artificial contours has little effect on intelligibility. Several objective measures were found to correlate well with intelligibility, and in particular the STOI and NFD measures. More detailed breakdown into gender, method of spectral smoothing and type of excitation shows the NFD to have consistently high correlation ($|r| \geq 0.77$) with subjective intelligibility.

6. Acknowledgements

We wish to thank Prof. J.M. Kates for supplying code for the HASQI and HASPI methods and the UK Home Office for supporting this work.

7. References

[1] L. Girin, J.-L. Schwartz, and G. Fang, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, Jun. 2001.

[2] A. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

[3] T. Baer, B. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times," *Journal of Rehabilitation Research and Development*, vol. 30, pp. 49–49, 1993.

[4] J. Kates, "An auditory model for intelligibility and quality predictions," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. Acoustical Society of America, 2013.

[5] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[6] ITU-T, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*. ITU-T recommendation P.862, 2000.

[7] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[8] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech communication algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, Jul. 2007.

[9] ——, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[10] A. Kressner, D. Anderson, and C. Rozell, "Evaluating the generalisation of the Hearing Aid Speech Quality Index HASQI," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 407–415, Feb. 2013.

[11] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 150, no. 5, pp. 2421–2424, Nov. 2006.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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 Communication*, vol. 27, pp. 187–207, Apr. 1999.

[13] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007," in *Proc. Blizzard Challenge 2007*, Aug. 2007.

[14] R. Rix, J. Beerands, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

[15] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*. Prentice-Hall, 1988.

[16] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, pp. 262–273, 1988.

[17] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[18] J. Kates, "On using coherence to measure distortion in hearing aids," in *Proceedings of Meetings on Acoustics*, vol. 91. Acoustical Society of America, 2013, pp. 2236–2244.

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

[20] J. Kates and K. Arehart, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[21] ——, "The Hearing-Aid Speech Quality Index (HASQI)," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117, may 2014.

[22] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*. Prentice Hall, 1989.

[23] ANSI, *Methods for calculation of the speech intelligibility index*. S3.5–1997 (American National Standards Institute, New York, 1997.

[24] R. A. Fox, E. Jacewicz, and C.-Y. Chang, "Auditory spectral integration in the perception of static vowels," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 6, pp. 1667–1681, 2011.