# Can Objective Measures Predict the Intelligibility of Modified HMM-based Synthetic Speech in Noise?

*Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King*

Centre for Speech Technology Research, University of Edinburgh, UK

C.Valentini-Botinhao@sms.ed.ac.uk, jyamagis@inf.ed.ac.uk, Simon.King@ed.ac.uk

## Abstract

Synthetic speech can be modified to improve intelligibility in noise. In order to perform modifications automatically, it would be useful to have an objective measure that could predict the intelligibility of modified synthetic speech for human listeners. We analysed the impact on intelligibility – and on how well objective measures predict it – when we separately modify speaking rate, fundamental frequency, line spectral pairs and spectral peaks. Shifting LSPs can increase intelligibility for human listeners; other modifications had weaker effects. Among the objective measures we evaluated, the Dau model and the Glimpse proportion were the best predictors of human performance.

**Index Terms**: objective measures for speech intelligibility, HMM-based speech synthesis, Lombard speech

## 1. Introduction

Objective measures are an essential tool for predicting quantities such as quality and intelligibility, that otherwise would have been obtained using subjective listening tests. If we were able to predict quality or intelligibility from only the acoustic signal, without the need for listening tests, we could use this prediction as a control mechanism. It could act, for instance, to control the effect of speech enhancement algorithms by minimizing the generated audible distortions, or it could control speech modifications designed to enhance certain acoustic properties of clean speech. Before including this control feedback into such algorithms, we need to make sure that the predictions made by objective measures remain accurate when applied to speech signals that have been modified by such enhancement techniques.

In previous studies we showed that objective measures based on models of the human auditory system are able to predict the intelligibility of HMM-based synthetic speech in diverse noisy situations [1]. HMM-based speech synthesis offers a versatile framework for modifying generated speech in order to increase intelligibility. Looking at natural speech produced in noise, sometimes known as "Lombard speech", can give clues as to how one could obtain an intelligibility gain. However too little is known about how the different acoustic modifications observed in Lombard speech contribute to this gain, and how that relates to the characteristics of the background noise.

Objective measures of intelligibility have already been evaluated when the mixture of noise and speech is processed through speech enhancement algorithms [2]. However thus far no study has been performed to show how the measures behave on clean speech that has been modified.

In this paper, we evaluate the impact on intelligibility of the following modifications applied to synthetic speech: changes in speaking rate, changes in fundamental frequency, shift of LSPs and enhancement of spectral peaks. We then evaluate several objective measures with regard to intelligibility prediction of this material. Four of the measures we evaluated were specifically designed to predict intelligibility – the Dau measure, the Glimpse proportion, the Short Time Objective Intelligibility (STOI) measure and the Speech Intelligibility Index (SII) – and one of them was specifically designed to measure quality – Perceptual Evaluation of Speech Quality (PESQ).

## 2. Hidden Markov Model-based Speech Synthesis

Hidden Markov Models (HMM) speech synthesis systems generate speech by using HMMs to model vocoder parameters [3]. The models are trained with parameters extracted from natural speech, to maximize the likelihood of the training data. The source can be represented by the fundamental frequency and aperiodic energy bands and the spectral envelope by Mel Generalized Cepstrum Line Spectral Pairs (MGC-LSP).

The statistical nature of HMM-based speech synthesis offers a great degree of control over the generated speech. By modifying the models of certain parameters we are able to control the acoustic characteristics of the generated speech without the need for new data.

The intelligibility of HMM-generated synthetic speech is comparable to natural speech in clean situations [4]. However synthetic speech can be made to be more intelligible than natural speech in noisy situations [5].

## 3. Objective Measures

There are many approaches to predict subjective dimensions of speech from the acoustic signal. The first approaches that were proposed were based on simple measures calculated on the spectral envelope, which is derived from linear predictive analysis. These conventional methods include the Cepstral Distance Measure (CEP), Log Spectral Distance (LSD), Itakura-Saito (IS) and Log-Likelihood Ratio (LLR) [6].

Following from the conventional methods are those measures that include some sort of frequency-dependent weighting inspired by psychoacoustics. Such measures include the Frequency Weighted Segmental SNR (FWS) [7] and the Weighted-Spectral Slope Metric (WSS) [8].

Incorporating those psychoacoustic findings, standards have been proposed to predict quality and intelligibility. The Perceptual Evaluation of Speech Quality (PESQ) [9] was designed as a measure for predicting the quality of speech signals transmitted over a telephone line. This measure includes an auditory transform and considers masking phenomena as well. The Speech Intelligibility Index (SII) [10] calculates a weighted SNR in the frequency domain, considering frequency-domain

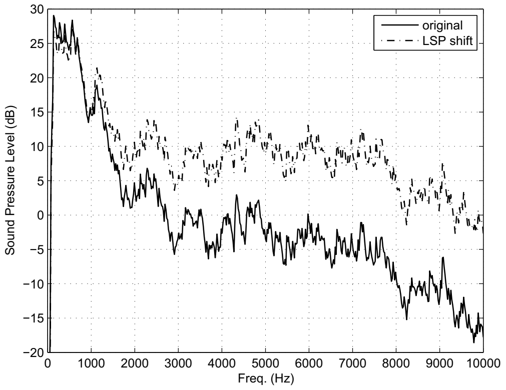


Figure 1: *Smoothed long term average spectrum of a non-modified speech sample and another sample in which the LSPs have been shifted towards higher frequencies.*

masking effects and auditory thresholds.

The objective measures of intelligibility that have been shown to best correlate with subjective scores for intelligibility tend to be ones that include elaborate auditory processing stages [2]. These measures compare an internal representation of the clean reference speech signal with an internal representation of the noisy signal, or of the noise alone, in order to predict how intelligible the noisy signal is.

In this group, notable measures include the Dau measure (DAU) [11], based on the Dau model [12] of the effective processing which takes place in the human auditory system. The model gives a time-domain representation that incorporates aspects of temporal adaptation. The measure is effectively the normalized correlation coefficient of the internal representation derived by the Dau model for both reference and noisy signal.

The Glimpse proportion measure (GP) [13] is derived from the Glimpse model for auditory processing. The measure is the proportion of spectral-temporal regions where speech is more energetic than noise, based on the idea that humans only attend to those ‘glimpses’ of speech that are not masked by noise.

The Short-Time Objective Intelligibility (STOI) [14] is the average linear correlation coefficient between a time-frequency representation of clean and noisy speech over time frames. This measure was proposed to work especially well for when noisy speech is processed by a time-frequency weighting algorithm for noise reduction or speech separation.

## 4. Listening Tests

The experimental strategy we adopted was to obtain subjective intelligibility scores using listening tests, for a wide range of noise types and SNRs. These scores were then correlated with the predictions made by various objective measures. In this section we explain the speech material that we used: the modifications we applied and the different types of listening situations, and the listening setup.

### 4.1. Speech Material and Modifications

In total, 96 different sentences were synthesized using an HMM-based Speech Synthesis System (HTS). The synthesis models were trained with 4000 sentences from a professional male British English speaker. We used 45 dimension mel-generalized cepstrum line spectral pairs (MGC-LSP) acoustic features to represent the spectral envelope. The training data waveforms were sampled at 48 kHz. The synthesized speech was produced at 48 kHz then downsampled to 20 kHz. The format of the test sentences was “name verb numeral adjective noun” (i.e., matrix sentences), with each word being chosen from a ten-word list.

In order to reproduce some of the effects found in natural Lombard speech we applied the following individual modifications to the synthesized material:

- spectral peak enhancement as described in [15]
- changes in the fundamental frequency: low / high
- shift of Line Spectral Pairs (LSPs) as described in [16]
- changes in the speaking rate as described in [17]: slow / fast

The strength of each modification was adjusted in such a way that it generated audible differences to the clean speech condition, but not necessarily intelligibility impacts in the noisy conditions. As an example, the effect of shifting the LSPs is shown in Fig. 1. In this figure we see the long term average spectrum of the original and modified speech signal when the LSPs are shifted towards the higher frequencies.

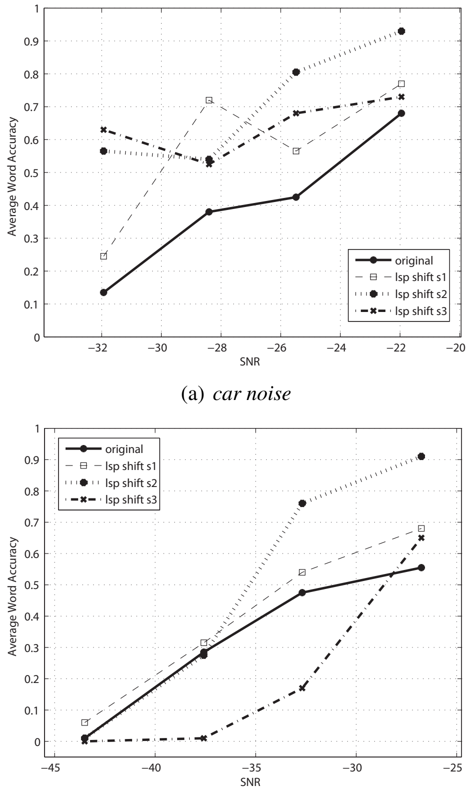
We can see that the spectral tilt becomes more flat and we expect that the formant frequencies also increase. For this experiment we shifted the LSPs towards the higher frequencies in three different steps, named here as s1, s2 and s3. The average spectral tilt of the non modified speech was $-1.51$ dB per octave and for the three strengths of shifts, s1, s2 and s3, the average spectral tilt becomes $-1.02$ dB, $-0.69$ dB and $-0.09$ dB per octave, respectively.

We used four different types of noise: speech shaped, cafeteria, car and high frequency noise, described in [1], at four different Signal to Noise Ratios (SNRs). The SNRs for each noise type were selected in a small calibration experiment performed with 9 participants. They correspond roughly to word accuracies of 0.2, 0.4, 0.6 and 0.8 obtained for each noise type when using non-modified speech sentences. The non-modified and the modified speech signals were first normalized sentence by sentence to yield the same overall signal level (rms) and then added to noise at those SNRs. That means that any intelligibility change observed when applying a certain modification is not the result of changes in overall energy levels. In total we generated 196 distinct listening situations from all combinations of speech modification type, modification strength (including a non-modified version), noise type and SNR.

### 4.2. Listening Setup

A total of 88 native English speakers with no reported hearing impairment participated in the listening experiment. Each participant listened to one quarter of all possible listening situations twice, each time with different sentences – a total of 96 sentences. The order of sentences and listening situations was random. The selection of listening situations from all possible situations was also random.

All signals were played at 20 kHz over headphones to participants in soundproof booths. Each individual sentence could be played only once before the participant had to type in what he or she heard. Each participant heard several examples that included different listening situations with modified and non-modified speech to familiarize themselves with the task before the actual experiment took place.



(a) *car noise*

(b) *high frequency noise*

Figure 2: *Average word accuracy by listeners for the 'shift LSP' condition, in car noise (top) and high frequency noise (bottom).*

# 5. Results and Discussions

In this section we report the results of the listening test, including the subjective scores and how well the objective measures correlate with them, for the different types of speech modification and listening situation.

## 5.1. Subjective Scores

We calculated the subjective score of word accuracy as the percent of correct words in a sentence, taking into account misspelling and spelling variations.

The only modifications that gave significant differences in intelligibility with respect to the original speech, across all noise types and SNRs, were the shifting of LSPs and the faster speaking rate. Slowing the speaking rate produced significant improvements in the presence of babble noise, and at some SNRs for speech-shaped noise. Lowering the fundamental frequency gave significant improvements only for high frequency noise at the highest SNR.

The largest improvements in average word accuracy were in the presence of car noise, shown in Fig. 2(a). For the lowest SNR case there was an improvement of 0.13 to 0.61 and for higher SNRs in that same noise condition the word accuracy improved from 0.38 to 0.72 and from 0.42 to 0.8. For the highest SNR level there was no significant improvement.

Shifting the LSPs does not always increase intelligibility though. For high frequency noise, as we can see in Fig. 2(b), a large shift in the LSPs (s3) results in a significant drop in word accuracy, while small shifts (s1,s2) give significant improvements.

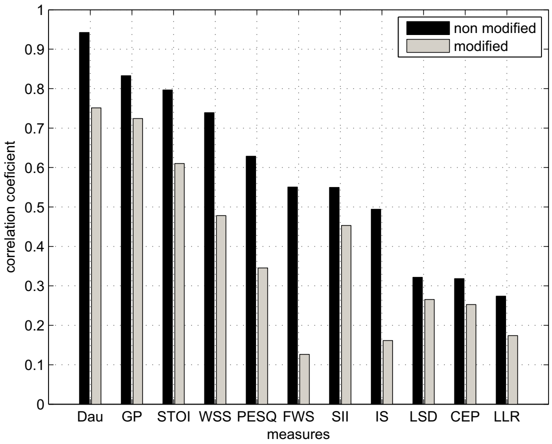


Figure 3: *Correlation coefficient between the subjective score and each objective measure, comparing performance on non-modified and modified speech.*

This result suggests that there is some optimal value of modification strength, and that this depends not only on the noise type, i.e. its spectral and temporal characteristics, but also on the SNR, i.e. on the noise energy level. Although it is found in natural Lombard speech, increasing fundamental frequency did not seem to provide any significant gains in intelligibility. Our result is consistent with another study, in which natural speech was modified [18].

## 5.2. Objective Measures

We compared the performance of each measure by extracting the normalized correlation coefficient $\rho$ using the subjective score for each listening situation group, averaged across listeners and sentences. These subjective scores were compared to the objective scores in the following manner:

$$\rho_i = \frac{\sum_{n=1}^{N}(S_n - \bar{S})(M_{i,n} - \bar{M}_i)}{\sqrt{\sum_{n=1}^{N}(S_n - \bar{S})^2 \sum_{n=1}^{N}(M_{i,n} - \bar{M}_i)^2}} \qquad (1)$$

where $S_n$ is the subjective score for listening situation $n$, $\bar{S}$ is the average score obtained for all situations in that group, $M_{i,n}$ is the objective score given by measure $i$ for listening situation $n$, $\bar{M}_i$ is the average score given by measure $i$ for all situations in that group.

To account for the non-linear relationship between subjective and objective scores we applied a logistic function to all objective measures and then calculated the correlation coefficient. The correlation coefficients for each objective measure obtained for the non-modified and modified speech are shown in Fig. 3; Fig. 4 shows them for each type of modification. On both figures the measures are ordered from left to right according to the correlation coefficient obtained for non-modified speech.

The results for non-modified speech displayed in Fig. 3 show that the measures based on auditory models outperform the other (conventional) methods. The conventional measures IS, LSD, CEP and LLR have no significant correlation with the subjective data: they do not predict intelligibility. The DAU measure outperforms all measures, obtaining a $\rho$ of 0.94, followed by the GP measure with 0.83 and STOI with 0.79.

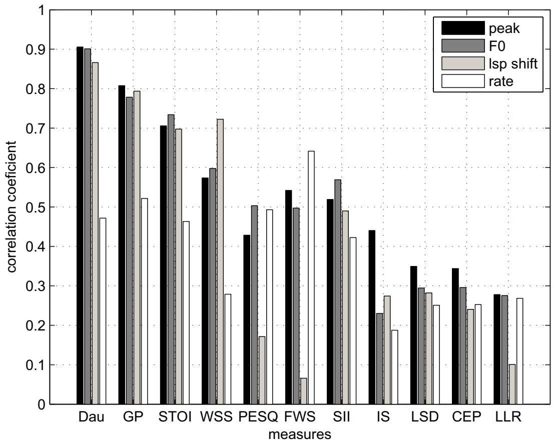


Figure 4: *Correlation coefficients obtained by the evaluated measures for each modification type.*

There is an overall drop in performance for all measures when predicting intelligibility of modified speech, especially for the FWS and IS measures. The DAU measure still obtains the highest correlation coefficient of $0.75$, followed by the GP measure with $0.72$ and STOI with $0.61$.

Analysing Fig. 4, we can see which modifications are responsible for this drop in performance. Most measures have a substantial loss in performance when speaking rate is altered, this particularly applies to the DAU, GP and STOI measures. The DAU and GP performance for other modification types remains similar to the performance obtained with non-modified speech, which indicates that the models underlying those measures are able to track the impact on intelligibility when altering speech in those ways, but are not necessarily good models for predicting the impact of changing speaking rate. One possible explanation is that changes in duration affect higher levels of processing involved in the listener's perception of speech; these cognitive levels are not taken into account by measures that attempt to model the auditory system only.

## 6. Conclusions

We evaluated how well various objective measures can predict the intelligibility of modified synthetic speech. By separately altering acoustic properties including fundamental frequency, speaking rate, LSP distribution and peaks in the spectrum envelope we were able to identify which sort of modifications have a significant effect on the intelligibility of synthetic speech in noise. The subjective intelligibility scores indicate that changes in fundamental frequency and spectral peaks do not alter intelligibility for the noise types we chose. Changes in the LSP distribution that flatten the spectral tilt tend to generate substantial intelligibility gains, depending on the noise type and SNR. Those objective measures that are based on auditory models – including the Dau and the Glimpse proportion measure – obtained correlation coefficients around or higher than $0.8$, except when dealing with changes in the speaking rate, when the correlations went as low as $0.5$. This indicates that these measures are good intelligibility predictors for modifications that take place in the spectral domain, but not for changes in duration.

## 7. Acknowledgment

Heiga Zen from Toshiba Research Europe Limited provided the recordings of car noise. The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 213850, 213845, and 256230 (SCALE, EMIME, and LISTA).

## 8. References

[1] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise," in *Proc. ICASSP*, Prague, Czech Republic, May 2011.

[2] C. Taal, R. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech intelligibility prediction," in *Proc. Interspeech*, Brighton, U.K., Sept. 2009, pp. 1947–1950.

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

[4] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge Workshop*, vol. 5, Brisbane, Australia, Sept. 2008.

[5] S. King and V. Karaiskos, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, Kyoto, Japan, Sept. 2010.

[6] J. Gray, A. and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, Oct. 1976.

[7] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. ICASSP*, vol. 3, Oklahoma, USA, April 1978, pp. 586–590.

[8] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra:a first step," in *Proc. ICASSP*, vol. 7, Paris, France, May 1982, pp. 1278–1281.

[9] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, Salt Lake City, USA, May 2001, pp. 749–752.

[10] *ANSI S3.5-1997 Methods for the calculation of the speech intelligibility index*, American National Standards Std., 1997.

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

[12] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[13] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

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

[15] Z. Ling, Y. Wu, Y. Wang, L. Qin, and R. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Proc. Blizzard Challenge Workshop*, Pittsburgh, USA, Sept. 2006.

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

[17] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 29–32.

[18] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Comm.*, vol. 51, no. 12, pp. 1253–1262, 2009.