EVALUATION OF OBJECTIVE MEASURES FOR INTELLIGIBILITY PREDICTION OF HMM-BASED SYNTHETIC SPEECH IN NOISE

Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King

The Centre for Speech Technology Research
University of Edinburgh
Edinburgh EH8 9AB, UK

ABSTRACT

In this paper we evaluate four objective measures of speech with regards to intelligibility prediction of synthesized speech in diverse noisy situations. We evaluated three intelligibility measures, the Dau measure, the glimpse proportion and the Speech Intelligibility Index (SII) and a quality measure, the Perceptual Evaluation of Speech Quality (PESQ). For the generation of synthesized speech we used a state of the art HMM-based speech synthesis system. The noisy conditions comprised four additive noises. The measures were compared with subjective intelligibility scores obtained in listening tests. The results show the Dau and the glimpse measures to be the best predictors of intelligibility, with correlations of around 0.83 to subjective scores. All measures gave less accurate predictions of intelligibility for synthetic speech than have previously been found for natural speech; in particular the SII measure. In additional experiments, we processed the synthesized speech by an ideal binary mask before adding noise. The Glimpse measure gave the most accurate intelligibility predictions in this situation.

Index Terms— objective measures for speech intelligibility, HMM-based speech synthesis

1. INTRODUCTION

Subjective measures involving human subjects are currently the most accurate indicators of quality and intelligibility. Humans utilise and reconcile information ranging from the pitch and spectral envelope to prosodic, semantic and pragmatic levels. However subjective tests are typically time consuming, expensive and not always reproducible. Our particular interest is in incorporating intelligibility measures into the optimization of statistical parametric speech synthesis – something that is obviously not tractable with subjective measures.

Several measures for speech quality and intelligibility have been proposed. They operate in different manners by prioritizing certain dimensions of the speech signal that are intended to reflect the perceptual cues that humans attend to when evaluating quality or intelligibility.

Predicting quality using objective measures has seen more success than predicting intelligibility. One of the most commonly used objective measures for speech quality, the Perceptual Evaluation of Speech Quality (PESQ) shows high correlation with Mean Opinion Score (MOS) for various types of distortions [1]. Speech measures for intelligibility – mostly based on the Speech Transmission Index (STI) [2] – do not correlate as well to subjective intelligibility scores.

There is still not a clear relationship between speech quality and intelligibility. There have been various studies evaluating speech quality measures as predictors of intelligibility. One of the most recent [3] compared conventional methods based on Signal to Noise Ratio (SNR) and LP coefficients (LPC) to perceptually-based measures and concluded that the latter are better predictors.

Several other studies have shown the correlation between subjective and objective measures for quality [4] and intelligibility prediction. However thus far no study has been conducted on how well objective measures correlate with subjective scores when the speech signal is generated by a text to speech system.

Natural and synthesized speech have different acoustic properties and prosody. Any of these differences could contribute to an intelligibility loss for synthesized speech. However, we do not know whether they will affect the performance of objective measures, which are mainly designed to with regard to the perceptually salient properties of *natural speech*.

In this paper, we evaluate four objective measures with regard to intelligibility prediction. Three of them were specifically designed to predict intelligibility – the Dau measure, the Glimpse proportion and the Speech Intelligibility Index (SII) – and the fourth measure was designed to measure quality – Perceptual Evaluation of Speech Quality (PESQ).

2. HMM-BASED SPEECH SYNTHESIS

HMM-based speech systems use statistical models, in this case Hidden Markov Models (HMM), in order to generate speech [5]. The models generate vocoder parameters that are used to generate speech. They are trained with parameters ex-

tracted from natural speech to maximize the likelihood of the training data.

The system is also trained with linguistic and prosodic contexts contained in labels that describe the text. This information is used for building clustering trees for duration, fundamental frequency and the spectral parameters.

Due to its statistical nature HMM-based speech synthesis has many advantages over the waveform concatenation systems [5]. However the excessive averaging that occurs in the training phase often results in less natural sounding speech. Intelligibility of HMH generated synthesized speech is comparable to natural speech [6].

3. LISTENING TESTS

In order to obtain the subjective scores we needed for the evaluation, we performed listening tests covering a range of conditions of noise and speech modifications. In this section we explain the speech material that we used and each of those conditions.

3.1. Test Material

We used so-called matrix sentences of the form "name verb numeral adjective noun". Each word in the sentence is chosen from a ten-word list. In total, 108 sentences were synthesized using the 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 linear spectral pairs (MGC-LSP) acoustic features as spectral features. The training sentences were sampled at 48 kHz. The synthesized speech was produced at 48 kHz then downsampled to 20 kHz.

We used four different types of noise: speech shaped, cafeteria, car and high frequency noise. The Long Term Average Spectrum (LTAS) of cafeteria, car and high frequency noise can be seen in fig. 1. The LTAS for the speech shaped noise was made similar to the cafeteria noise. The speech shaped and the high frequency noises were generated from white noise. The cafeteria and car noises were actual recordings and are non stationary.

In total we created 36 different listening situations, by varying the noise and speech modification. The first set of situations, where no modification was applied to the speech, constitute 20 of these (four different additive noises added at five different levels of speech).

The second set employed modified speech and constitute the other 16 situations (four noises added at two different levels of speech to two different speech modifications).

The modified speech was created from clean speech by applying an Ideal Binary Mask (IBM). The mask is applied to clean speech before mixing it with noise in order to enhance those time frequency bins of speech that are higher than the noise while removing the bins that are not, with the aim of increasing intelligibility of the mixture. The threshold used to create the IBM was a parameter we varied, to create the two versions of modification.

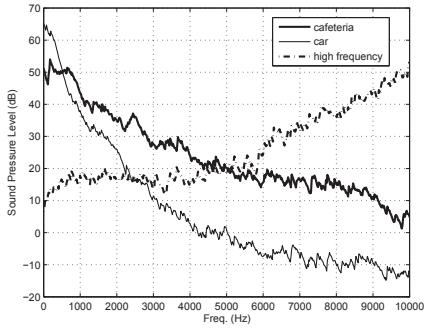

Fig. 1. Long Term Average Spectrum (LTAS) in sound pressure level of cafeteria, car and high frequency noise. The speech shaped noise LTAS was set to match the cafeteria LTAS

3.2. Listening Setup

A total of 41 native English speakers with no reported listening impairment participated in the listening experiment. Each participant listened to each situation three times with different sentences each time and in a random order. All signals were played at 20 kHz over headphones to participants in sound proof booths. Each individual sentence could be played only once before the participant had to type in what he or she heard.

4. OBJECTIVE MEASURES

The objective measures of intelligibility that best correlate with subjective scores tend to be ones that include elaborate auditory processing stages [3]. These measures compare an internal representation of the clean reference speech signal with an internal representation of the noisy signal, in order to predict how intelligible the noisy signal is.

The Dau measure [7] is based on the Dau model [8] of the processing that takes places in the human auditory system. The model is 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 correlation is taken over a 30 ms window frame, with a frame shift of $10ms$. The measure is then the averaged over the frames that present high energy levels [7].

The Glimpse measure [9] comes from the Glimpse model for auditory processing. The model is based on the assumption that in a noisy environment humans listen to the glimpses of speech that are less masked. The internal representation

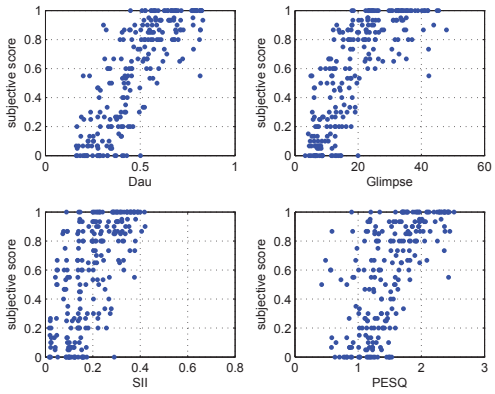

Fig. 2. Scatter plots relating subjective word accuracy scores for non modified speech to each objective measure (a) Dau (b) Glimpse (c) SII (d) PESQ

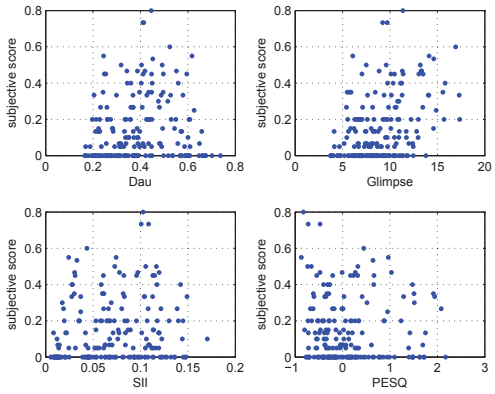

Fig. 3. Scatter plots relating subjective word accuracy scores for modified speech to each objective measure (a) Dau (b) Glimpse (c) SII (d) PESQ

	Dau	Glimpse	SII	PESQ
ρ	**0.80**	0.76	0.63	0.65
σ	0.25	21.20	0.38	0.99

Table 1. Root mean square error σ and normalized correlation coefficient ρ for each objective measure (before mapping) for non modified speech

	Dau	Glimpse	SII	PESQ
ρ	0.08	**0.42**	0.07	-0.05
σ	0.32	9.19	0.19	0.72

Table 3. Root mean square error σ and normalized correlation coefficient ρ for each objective measure (before mapping) for modified speech

	Dau	Glimpse	SII	PESQ
ρ	**0.83**	0.82	0.64	0.62
σ	0.22	0.23	0.28	0.30

Table 2. Root mean square error σ and normalized correlation coefficient ρ for each objective measure (after mapping) for non modified speech

	Dau	Glimpse	SII	PESQ
ρ	0.13	**0.42**	0.07	-0.01
σ	0.48	0.27	0.40	0.30

Table 4. Root mean square error σ and normalized correlation coefficient ρ for each objective measure (after mapping) for modified speech

is in time-frequency and is derived using Gammatone filter banks. The measure is the proportion of spectral-temporal regions where the speech is more energetic than the noise. Like the Dau measure, it is also calculated framewise .

The SII [10] calculates a weighted Signal to Noise Ratio (SNR) in the frequency domain, considering frequency domain masking effects and auditory thresholds. The sum of the weighted SNR produces the intelligibility estimate.

The Perceptual Evaluation of Speech Quality (PESQ) [1] is a measure designed for predicting the quality of speech signals transmitted over a telephone line. The measure includes an auditory transform and considers the masking phenomena for the comparison of this transformed representation. PESQ can not handle wideband speech signals because it was specially designed for narrowband signals.

We calculated the average subjective score as the percent of correct words in a sentence across all participants in the listening test. Each sentence/situation combination used in the listening test is shown as a point in each of the scatter plots in fig. 2. The plots show the relationship between the average subjective scores and each of the objective measures for the non modified speech condition. This relationship is especially non-linear for the Dau and Glimpse measures. The plots in fig. 3 show the relationship for the modified speech conditions. The relationships are a lot less obvious in most cases.

4.1. Improving the correlation by mapping

To improve the fit, we tried using a logistic function to map the values retrieved from each objective measure to the average subjective scores obtained in the listening test for each situation $M_i = \frac{1}{1+\exp^{(O-m_i)/S}}$ where offset O and slope S are the fitting parameters and m_i and M_i are the objective measure i before and after mapping. We used the average score obtained for each noisy/speech condition across all sentences

and listeners to find O and S, done separately for the modified and non-modified speech conditions.

5. RESULTS AND DISCUSSIONS

In order to evaluate the measures we extracted the normalized correlation coefficient ρ and the root mean square error σ. These error measures were calculated using the subjective scores given to each sentence/situation combination and averaged across listeners that heard the same combination. These subjective scores were compared to the (possibly mapped) objective scores in the following manner:

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

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(S_n - M_{i,n})^2} \quad (2)$$

where S_n is the subjective score for sentence/situation n, $\bar{S}$ is the average score obtained by all sentence/situation, $M_{i,n}$ is the objective score obtained by measure i for sentence/situation n, $\bar{M_i}$ is the average score obtained by measure i for all sentence/situation.

The pair of tables 1 & 3 show the evaluation measures for when speech was not modified and when it was, using linear regression. The pair of tables 2 & 4 show the same measures, after mapping using the logistic regression. For the condition where speech has not been modified, we can see that the Dau and Glimpse measures are the better predictors for intelligibility, with correlation coefficients of 0.83 and 0.82, and smaller root mean square errors (which can only be compared across measures for the mapped condition, i.e., tables 2 & 4). SII and PESQ obtained lower correlations of 0.64 and 0.62 and had larger errors.

When we compare the results in table 2 with correlation coefficients obtained in other studies [3, 9, 7] we observe a loss of prediction performance for all measures when the speech is synthetic rather than natural, particularly for the SII measure.

The results in table 4 show that all measures perform worse for *modified* synthetic speech. The Glimpse models seems to perform best, obtaining a correlation coefficient of 0.42 and the smallest errors. This result could be expected because this measure predicts intelligibility from the proportion of time-frequency bins that are above the noise, which matches the type of modification we performed.

6. CONCLUSIONS

We evaluated four different objective measures with regard to speech intelligibility prediction of mixtures of noise and synthesized speech. We found that Dau and Glimpse measures exhibited similar performance as did PESQ and SII, with the former proving to be the better intelligibility predictors. Overall, all measures seem to have a loss in performance when compared to predicting intelligibility of mixtures with natural speech. We aim to investigate this further. For speech processed with an ideal binary mask (intended to improve subjective intelligibility), the Glimpse model gave better predictions than the other measures. Future work will use a wider range of types of speech modification.

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).

7. REFERENCES

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

[2] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

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

[4] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

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

[6] 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 Text-to-Speech Workshop*, 2008, vol. 5.

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

[8] 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.

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

[10] "ANSI S3.5-1997 Methods for the calculation of the speech intelligibility index," 1997.