AN ASSESSMENT OF AUTOMATIC SPEECH RECOGNITION AS SPEECH INTELLIGIBILITY ESTIMATION IN THE CONTEXT OF ADDITIVE NOISE

Wei M. Liu, John S. D. Mason, Nicholas W. D. Evans and Keith A. Jellyman

School of Engineering, University of Wales Swansea
Singleton Park, Swansea, SA2 8PP, UK
{199997, j.s.d.mason, n.w.d.evans, 174869}@swansea.ac.uk

ABSTRACT

This paper investigates the potential applicability of automatic speech recognition (ASR) and 6 well-reported objective quality measures for the task of ranking intelligibility of speech degraded by different real life background noises. In a recent investigation ASR has been reported to give high subjective correlation with human assessment when tested with various system degradations. This paper extends this investigation in two directions. First, the usefulness of the measures in the context of different real-life noises is considered. Second, the direct correspondence between statistics computed by an ASR system and human perceived intelligibility is assessed. Subjective listening tests are carried out to provide ground truth. Results show that ASR and WSS (weighted spectral slope) are the only two measures out of the seven considered to give good correlation with human opinion. Specially noted is performance of ASR with correlations ranging from 0.77 to 0.90.

Index Terms: speech intelligibility, objective quality measures, ASR

1. INTRODUCTION

Speech intelligibility assessment is an important topic in communication. The loss of intelligibility simply means that communication does not exist. More emphasis has been placed on the more general overall quality rather than specifically on intelligibility due possibly to the explosion of commercial communication systems where overall quality is important. This is reflected in the relative lack of advances in the area of objective assessment specific to intelligibility.

The past 3 decades have witnessed significant research efforts directed to the area of overall quality assessment. The early work of Quackenbush et al [1] reported a thorough investigation of over 2000 variations of waveform-based and spectral-based objective quality measures, including signal-to-noise ratio (SNR), the Itakura-Saito (IS) distance, the log-likelihood ratio (LLR), the weighted spectral slope (WSS), the cepstral distance (CD) and among others. More recent developments have followed a perceptual-based approach. Explicit models for some of the known attributes of human auditory perception are incorporated into the quality assessors with the motivation to create assessors that better mimic the human hearing system. Such measures include the early Bark spectral distortion (BSD) proposed by Wang in 1992 [2], its improved version, modified BSD (MBSD) by Yang [3], Measuring Normalizing Blocks (MNB) by Voran [4], and perceptual evaluation of speech quality (PESQ) by Beerends et al [5]. All report good correlation with subjective results over a large range of degradations. Of particular note is PESQ [6] which was standardised as ITU-T Recommendation P.862 in 2002 and is widely acknowledged as the state-of-the-art.

As mentioned, development in the specific area of intelligibility assessment is relatively rather inactive compared to that of the more general quality assessment. Early attempts date back to 1947 when Bell Labs developed the articulation index (AI) [7]. However, progress has become somewhat stagnant since the development of speech transmission index (STI) by Houtgast and Steeneken [8] in 1973 which is included in IEC standard 60268-16. Subsequent works evolved mainly around enhancement or simplification of STI. Recently, both Chernick et al [9] and Jiang et al [10] investigated ASR in this context with the DoD-CELP and G.729 codec respectively. Promising results are reported. These findings in part provide the motivation for the work published recently by the current authors in ICASSP 2006 [11]. The paper reports on the potential of the same measures assessed here for the tasks of intelligibility assessment in the context of standard coding distortions with different system configuration. Experimental findings show that ASR emerges to be a reliable intelligibility estimator for the degradations considered.

This paper serves to extend the previously published work by (i) investigating the usefulness of the measures in the context of additive background noise, and (ii) investigating the correspondence between various ASR statistics and human perceived intelligibility. The types of degradations considered are 8 different real life noises. The paper is structured as follows: Section 2 briefly describes the objective measures; Section 3 describes the experimental works including subjective listening tests, followed by results and discussions in Section 4 and conclusions in Section 5.

2. OBJECTIVE MEASURES

Six of the measures considered span the evolution of objective quality assessments from waveform-based to perceptualy-based measures. All measures have been reported at one time or another to give high correlation with human perceived quality under a large variety of degradations. ASR has not been widely used in the context of quality/intelligibility assessment, however, investigations have indicated its potential [9–11]. Note that in this section all correlations quoted are correlations with quality unless stated otherwise. In subsequent sections the correlations are primarily with intelligibility.

2.1. Segmental Signal-to-Noise Ratio (SegSNR)

SegSNR is an improvement of the classical SNR. Signal-to-noise ratio (SNR) is determined from each frame after which an upper and lower threshold is set to replace frames with exceptionally high or low SNR. The quality estimate is then the average SNR from all frames. A correlation of 0.77 is reported by Quackenbush et al's [1] but its application is limited to testing of waveform coder distortions.

2.2. Cepstral Distance (CD)

CD is essentially the comparison of two smoothed spectra in the cepstral domain. Kitawaki [12] observed that spectral envelope measures correspond better to subjective results than whole spectral measures; CD achieved a correlation of 0.87 and is strongly proposed as an accurate quality estimator for low-bit rate coding systems and other non-linear distortions alike.

2.3. Weighted Spectral Slope (WSS)

WSS by Klatt [13] is based on weighted differences between the spectral slopes in each of 36 overlapping frequency bands with bandwidths analogous to that of critical bands. Its correlation at 0.74 is the highest from Quackenbush et al's [1] study. In the context of intelligibility, WSS scores well in [11] with a correlation of 0.83-0.87.

2.4. Measuring Normalising Blocks (MNB)

MNB was introduced by Voran [4] in 1995. It is somewhat distinctive from other perceptual-based measures in that it employs a simple perceptual transformation module. A sophisticated cognition module follows which consists of a hierarchy of measuring normalising blocks for emulating human patterns of adaptation and reaction to spectral deviations that span different time and frequency scales. In [4] this measure is reported to outperform CD, BSD and ITU-T Rec. P.861 (PSQM) with an average correlation coefficient of about 0.97 when tested on 219 different degradation conditions.

2.5. Modified Bark Spectral Distortion (MBSD)

MBSD [3] assumes that speech quality is directly related to speech loudness. The measure transforms energies to Bark frequency domain where the Bark coefficients are then transformed to dB to model perceived loudness. A masking threshold is incorporated where distortion below the threshold is excluded from the calculation. MBSD gives correlation coefficient at 0.96 when tested on a modulated noise reference unit(MNRU) and a large range of coding distortions.

2.6. Perceptual Evaluation of Speech Quality (PESQ)

PESQ [5] compares two perceptually-transformed signals and generates a noise disturbance value to estimate the perceived speech quality. It was standardised as ITU-T Recommendation P.862 in 2001 replacing PSQM (ITU-T Rec. P.861). PESQ has an improved time-alignment module which makes it more robust for use in real networks with varying delays. PESQ aims to give quality indications which mimic the Mean Opinion Score (MOS).

2.7. Automatic Speech Recognizer (ASR)

The motivations for the use of ASR include: (i) the observation that word recognition performed by ASR can be thought of as machine intelligibility; (ii) the recent positive findings of Chernick et al [9], Jiang et al [10] and Liu [11] that suggest a good correlation between human intelligibility and machine recognition.

3. EXPERIMENTS

The experiments presented here illustrate the potential of the 7 objective measures mentioned in Section 2. The performances of the objective measures were judged by correlations between their estimates and results from subjective listening tests.

3.1. Database

Both subjective and objective tests were conducted using the TIDigits database. The database has 11 words in its vocabulary, namely the digits one to nine, 'oh' and 'zero'. Though this database is not specially designed for intelligibility assessment, it was chosen here first as it provides a straightforward scoring process for subjective tests, with minimal influence from listeners' vocabulary power, and second because it is explicitly configured for ASR. The same database was used by Hicks et al [14] for ASR testing of speech intelligibility.

Degradations considered include 8 real life background noises used in the production of the Aurora 2 digit string corpus [15], namely airport, babble (crowd of people), car, exhibition hall, restaurant, street, subway (suburban train) and train station noises. They represent the most probable application scenarios for telecommunication terminals. 566 clean four-digits strings were selected from the TIDigits database and the noises were added to them using the standard noise addition software from ITU-T Rec, P.56 [16] at signal-to-noise ratio ranging from -10dB to 0dB at 0.5dB interval. In total there were 168 degradation conditions (8 noise types * 21 SNRs).

3.2. Subjective Tests

Human listening tests were carried out to collect the ground truth for comparison with objective results. The interface of a system designed to assist in the task is shown in Figure 1. The test required the listeners to key in digits heard. The '?' key was hit when a digit is incomprehensible while the '????' key (equivalent to hitting the '?' key 4 times) was for when the whole utterance is incomprehensible.

The subjective tests involved 5 untrained (naive) human subjects aged 24 to 55 with healthy listening ability. Every subject repeated 5 tests for each noise type, totalling 40 tests per subject. One test consists of 21 test signals corrupted by the same noise type at SNRs ranging from -10dB to 0dB with 0.5dB interval, i.e. one test signal per SNR. The test signals were randomly chosen from respective complete testsets. There was no repeating of test utterances either within one test or across tests of the same noise type in order to avoid the human subject from memorizing the test signals.

Three indicators have been devised to quantify the level of intelligibility as indicated by the human subjects. The first indicator, $SUBJcorrect$ is simply the number of digits identified correctly. The second indicator, $SUBJmiss$ is defined as digits lost regardless of its position in the test utterance. For instance, 3261 heard when 3615 is played would incur only one $SUBJmiss$ score. Lastly, $SUBJdunno$ is the number of unrecognisable digits, i.e. the number of '?' responses given. Both $SUBJmiss$ and $SUBJdunno$ are recognition errors and are inverted to indicate intelligibility.

3.3. Objective Assessments

The 7 objective measures considered were SegSNR, CD, WSS, MNB, MBSD, PESQ and ASR. The first six measures are based on an intrusive approach in that a reference signal is needed in order to compute intelligibility difference between the reference and test signal. References used were the corresponding clean, unprocessed signals. ASR does not require a corresponding reference signal. Instead a set of 8440 clean utterances were used to train the recogniser.

All measures with exception of ASR give one single indication in terms of either quantity of distortion or quality (assumed as intelligibility in the context of this study). The ASR used here provides two main indications, namely word accuracy, $ASRacc$, and percent-

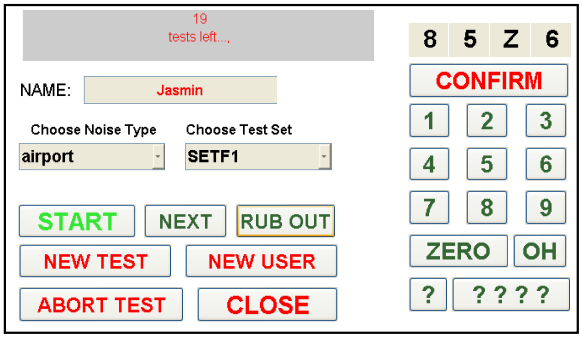

Fig. 1. Graphical user interface designed for subjective listening tests.

	SUBJcorrect	SUBJdunno	SUBJmiss
SegSNR	0.18	0.20	0.22
CD	-0.40	-0.41	-0.47
WSS	0.55	0.58	0.56
MNB	-0.26	-0.40	-0.33
MBSD	-0.64	-0.66	-0.62
PESQ	-0.52	-0.62	-0.59
ASRacc	0.89	0.85	0.85
ASR%correct	-0.46	-0.54	-0.52
ASRdel	-0.83	-0.86	-0.83
ASRsubst	0.88	0.90	0.86
ASRins	0.77	0.82	0.80

Table 1. Correlation coefficients of 7 different objective measures using 3 different subjective indicators.

age correct, $ASR\%correct$

$$ASR\%correct = \frac{total - ASRsubst - ASRdel}{total} \quad (1)$$

$$ASRacc = \frac{total - ASRsubst - ASRdel - ASRins}{total} \quad (2)$$

where $total$ is the total digits under test and $ASRins$, $ASRsubst$ and $ASRdel$ are insertion, substitution and deletion respectively according to usual ASR terminology. All five ASR statistics ($ASRacc$, $ASR\%correct$, $ASRsubst$, $ASRins$ and $ASRdel$) are investigated for their correlation with the ground truth.

Intelligibility associated with a particular noise type is the mean score across all 566 signals for every SNR averaged across all SNRs considered, i.e. -10dB to 0dB. Those results given in terms of distortion indication (WSS, CD, LLR, MBSD, $ASRsubst$, $ASRins$ and $ASRdel$) were inverted to indicate level of intelligibility.

4. RESULTS AND DISCUSSIONS

Performances of the objective measures are presented in terms of the Pearson product-moment correlation coefficient, r,

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y} \quad (3)$$

where X and Y are the subjective and objective scores, with means $\overline{X}$ and $\overline{Y}$, and standard deviation S_X and S_Y respectively, while n is the number of degradations considered. The coefficient ranges from -1 to 1 with 1 being the highest-correlated to subjective scores and vice-versa. Table 1 show the correlations computed using three different subjective indicators.

Several observations can be made from Table 1. Primarily, ASR and WSS appear to be the only measures that show reasonable positive correlations with human opinions. Three ASR statistics ($ASRacc$, $ASRsubst$ and $ASRins$) are the highest correlated with correlation at 0.86, 0.88 and 0.80 respectively when averaged across the 3 columns. This is followed by WSS at 0.56. Most other measures show negative correlations, demonstrating the outstanding potential of WSS and ASR in predicting intelligibility ranking across signals corrupted by different background noises. This is further illustrated in Figure 2 where side-by-side comparisons can be made. The noise types on the x-axis are ordered according to the subjective scores: babble noise (left most) is associated with the lowest human intelligibility score while subway noise (right most) is the highest. Figure 3 plots the same for scores of MNB, MBSD, and PESQ. Most measures erroneously indicate that noises with speech-like features (e.g. babble and restaurant noise) cause less damage to speech intelligibility than other fairly stationary noises. However, human opinions indicate otherwise. Notice that bars for both human scores and objective measure scores follow similar trend in Figure 2. However, no such trend exists in Figure 3.

The second observation is that not all ASR statistics are directly relevant to intelligibiliy. While $ASRacc$, $ASRsubst$ and $ASRins$ correlate well with subjective results, $ASRdel$ and $ASR\%correct$ show negative correlations. Several reasons might be postulated: (i) $ASRacc$ correlates well especially with $SUBJcorrect$ because both are computed using essentially the same approach, i.e. total tests minus all possible errors: possible errors in ASR being insertion, deletion and substitution; in listening test being number of wrong answer and '?' response. (ii) The recognizer is prone to mistaking noise as speech when tested with speech-like noises such as babble. As a result the occurance of insertion and substitution increase while deletion decreases. Hence perhaps predictably $ASRdel$ is not a useful indicator as less deletion does not imply higher intelligibility in this context. However, $ASRsubst$ and $ASRins$ are useful as they are able to identify speech-like noises that cause greater impairment to intelligibility. The absence of insertion as a useful indicator also explains the bad correlation given by $ASR\%correct$.

The third observation is that all three subjective indicators agree with each other which indirectly confirms the reliability of the listening tests. Certain indicators however correlate better with certain objective measures. The best match is perhaps $SUBJdunno$-$ASRsubst$ with correlation at 0.90 due, conceivably, to the direct correspondence between the definitions of the two variables. Both refer to failure to recognize the words despite knowing the existence of the words in the test signals, as opposed to $ASRdel$ which refers to failure to even detect the existence of the words.

One limitation of this preliminary work is that it has been designed for fixed-length test signals. As a result insertion and word accuracy cannot be identified from the subjective tests since the human subjects know how many digits are to be played. Immediate further work is therefore to repeat the tests using signals of variable length to further investigate any correlation between transcriptions generated by the ASR and by human.

Overall, the results raise concerns about the poor performance of the quality measures when applied specifically to intelligibility. It is speculated that in the region of the intelligibility threshold, the other quality components such as loudness, naturalness and ease of listening are so low that the intelligibility component is totally swamped.

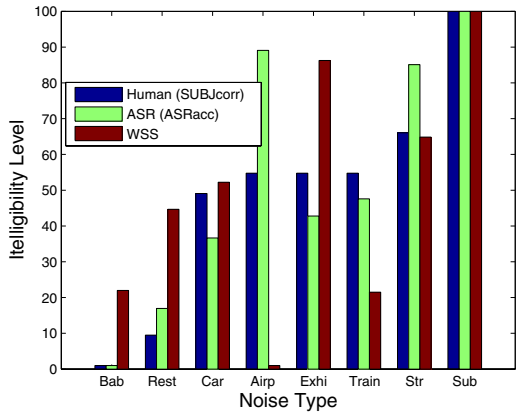

Fig. 2. Normalised WSS and ASRacc scores plotted against subjective scores using SUBJ%correct. x-axis: [Babble, Restaurant, Car, Airport, Exhibition hall, Train station, Street, Subway].

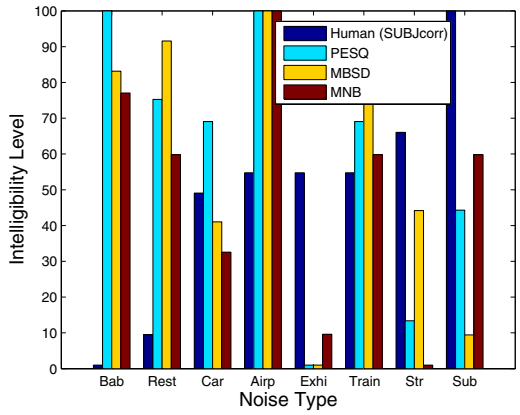

Fig. 3. Normalised PESQ, MBSD and MNB scores plotted against subjective scores using SUBJ%correct.

The potential of ASR is thus an important finding that warrants further investigations.

5. CONCLUSION

ASR and six different widely-reported quality measures are assessed for their applicability in the estimation of speech intelligibility in the context of additive background noise. Results show that WSS with average subjective correlation at 0.56 could potentially be used in estimating intelligibility, however most other quality measures including the state-of-art ones such as PESQ perform comparatively poorly. On the other hand, ASR stands out as the best measure among the seven studied here with correlation as high as 0.90. Though not all ASR statistics prove to be useful, it is most positive to observe that most agree with similar statistics as perceived by humans. The potential of ASR needs to be further investigated under contexts other than additive background noise.

6. REFERENCES

[1] S.R. Quackenbush, T.P. Barnwell III, and M.A. Clement, "Objective Measures of Speech Quality," Prentice Hall, Englewood Cliffs, 1988.

[2] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," IEEE J. Select. Areas Comm., vol. 10, pp. 819-829, June 1992.

[3] W. Yang, M. Benbouchta, and R. Yantorno, "A Modified Bark Spectral Distortion Measure as an Objective Speech Quality Measure," IEEE ICASSP , pp. 541-544, 1998.

[4] S. Voran, "Estimation of Perceived Speech Quality using Measuring Normalizing Blocks," IEEE Speech Coding Workshop, pp. 83-84, 1997.

[5] J.G. Beerends, A.W. Rix, M.P. Hollier, and A.P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment, Part I . Time-Delay Compensation," J. Audio Eng. Soc., Vol. 50, No. 10, 2002.

[6] ITU-T Recommendation P.862 "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", February 2001.

[7] K.D. Kryter, "Methods for the Calculation and Use of the Articulation Index," J. Acoust. Soc. Am., Vol. 34, pp. 1689-1697, 1962.

[8] H.J.M. Steeneken, T. Houtgast, "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," Acustica 28, pp. 66-73, 1973.

[9] C.M. Chernick, S. Leigh, K.L. Mills, and R. Toense, "Testing the Ability of Speech Reconizers to Measure the Effectiveness of Encoding Algorithms for Digital Speech Transmission," IEEE Int. Military Comm. Conf. (MILCOM), 1999.

[10] W. Jiang, H. Schulzrinne, "Speech Recognition Performance as an Effective Perceived Quality Predictor," IEEE Int. Workshop on Quality of Service, pp. 269-275, 2002.

[11] W.M. Liu, K.A. Jellyman, J.S.D Mason, and N.W.D. Evans, "Assessment of Objective Measures for Speech Intelligibility Estimation," to appear in ICASSP, 2006.

[12] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective Quality Evaluation for Low-Bit-Rate Speech Coding Systems," IEEE J. on Sel. Areas in Comm, pp. 242-248, 1988.

[13] D.H. Klatt, "Prediction of Perceived Phonetic Distance from Critical-band Spectra: A First Step," IEEE ICASSP, pp. 1278-1281, 1982.

[14] W.T. Hicks, B.Y. Smolenski, and R.E. Yantorno, "Testing the Intelligibility of Corrupted Speech with an Automated Speech Recognition System," 7th World Multiconference on Systemics, Cybernetics and Informatics, 2003.

[15] H.G. Hirsch, D. Pearce, "The Aurora Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR 2000, Sept 2000.

[16] ITU recommendation P.56, "Objective Measurement of Active Speech Level", Mar. 1993.