DATA DRIVEN METHOD FOR NON-INTRUSIVE SPEECH INTELLIGIBILITY ESTIMATION

Dushyant Sharma[1], Gaston Hilkhuysen[2], Nikolay D. Gaubitch[1],
Patrick A. Naylor[1],Mike Brookes[1],Mark Huckvale[2]

Centre for Law Enforcement Audio Research (CLEAR)

[1] Electrical and Electronic Engineering,
Imperial College London, UK
email: {dushyant.sharma02, ndg, p.naylor,
mike.brookes}@ic.ac.uk

[2] Speech, Hearing & Phonetic Sciences,
University College London, UK
email: {g.hilkhuysen, m.huckvale}@ucl.ac.uk

ABSTRACT

We propose a data driven, non-intrusive method for speech intelligibility estimation. We begin with a large set of speech signal specific features and use a dimensionality reduction approach based on correlation and principal component analysis to find the most relevant features for intelligibility prediction. These are then used to train a Gaussian mixture model from which the intelligibility of unseen data is inferred. Experimental results show that our method gives a correlation with subjective intelligibility of 0.92 and a correlation of 0.96 with the ANSI standard Speech Intelligibility Index.

1. INTRODUCTION

Speech intelligibility is a measure of how much of what is spoken is recognized by a listener. It is an important quantifier for speech communication applications in telecommunications, hearing aids and intelligence gathering in law enforcement applications. Intelligibility scores can be classified as being either subjective or objective.

Subjective speech intelligibility scores are obtained through listening experiments where subjects listen to speech samples and are either asked to repeat the words they have heard or else to select one from a predefined set of answers. It is necessary to perform the experiments on many subjects in order to get a statistically reliable estimate, which makes the task of obtaining subjective intelligibility scores expensive and time consuming. Objective intelligibility estimation that can be performed algorithmically is clearly advantageous and several methods have been developed, including, the ANSI standard Speech Intelligibility Index (SII) [1] that is a development of the Articulation Index (AI) [2]. These measures are intrusive in nature as they require knowledge of the clean speech signal, and although they are useful in controlled experiments, there are many situations where only the noisy speech signal is available; in such cases, it would be valuable to have a non-intrusive measure that operates directly on the observed signals.

We propose a data driven approach to non-intrusive intelligibility estimation inspired by the Low Complexity Speech Quality Assessment (LCQA) method developed by Grancharov *et al.* [3]. We begin by defining a large set of local and global speech specific features. Subsequently, we employ a dimensionality reduction scheme based on correlation and Principal Component Analysis (PCA) in order to find the features that are best suited for predicting speech intelligibility. Finally, these features are used to train a Gaussian Mixture Model (GMM) which is used to infer the intelligibility of new, unseen, data from the noisy speech signal alone.

The remainder of the paper is organized as follows. In Section 2, we review the LCQA method as it was originally proposed, for non-intrusive quality estimation. We then show, in Section 3, how the LCQA framework can be developed for our non-intrusive intelligibility measure. Section 4 presents results of our measure in terms of its correlation with subjective intelligibility scores as well as with intrusive intelligibility measures. Finally, conclusions from this work are drawn in Section 5.

2. LCQA REVIEW

LCQA [3] is a data driven approach to speech quality evaluation which has been shown to correlate well with subjective Mean Opinion Score (MOS) [4]; the correlation is higher than that of the standard ITU-T P.563 [5] which, like LCQA, is non-intrusive. In the following, we summarize the key features of LCQA and refer the reader to [3] for further details.

A frame selection scheme is developed using thresholds applied to the spectral flatness, spectral dynamics and the speech variance per frame features. This allows a flexible voice activity detection to be performed, based on the optimization of the feature thresholds that maximize the quality estimate. The algorithm models the statistical properties of the per frame features using their mean, variance, skewness and kurtosis. In modeling the global properties of the optimal per frame features, the dimensionality of the feature space is significantly reduced to 44 features for each speech utterance.

In order to optimize the performance of the classification, it is required to retain the minimum number of global features that maximize the estimation criteria (quality in the original context). This is achieved by the sequential floating backward selection algorithm [6], [7]. After a minimization of the root-mean-square error (RMSE) performance of the LCQA algorithm, the final feature vector is reduced to 14 dimensions.

The LCQA algorithm is trained on a large number of speech utterances (typically 2 sentences separated by a small pause) that have been subjectively labeled (through listening experiments for example) with the mean opinion score (MOS) [8]. Fourteen global features are extracted for each utterance and a GMM is trained on the joint distribution of

the global features and the MOS for each utterance. The GMM containing M mixtures is defined by a set of mean vectors, covariance matrices and mixture weights, estimated using the Expectation Maximization (EM) algorithm [9].

The global feature vector describes the statistical properties of certain aspects of the speech signal; at no point explicit auditory or cognitive modeling is performed. This suggests that the algorithm framework may be able to model different subjective criteria, such as the intelligibility of the utterance.

3. NON-INTRUSIVE INTELLIGIBILITY ASSESSMENT

In this section, we describe the Low Cost Intelligibility Assessment (LCIA) algorithm for estimating the speech intelligibility by deriving per frame features from the speech waveform, then applying a statistical model followed by a dimensionality reduction and GMM mapping. We also describe the database used for evaluation of the algorithm and the training procedure.

3.1 Algorithm overview

The key algorithm blocks are illustrated in Fig. 2 and described further in this section.

3.1.1 Derived Features

The first step is a Linear Prediction Coding (LPC) using 20 ms, non overlapping windows of the speech signal. The frequency response of the LPC coefficients is used to derive a number of per frame features including the spectral flatness, spectral centroid, excitation variance and spectral dynamics. In addition, the speech variance and the *iSNR* (defined in Section 3.2) per frame are computed giving a total of 6 per frame features. In addition, the first time derivatives of these (except spectral dynamics) are also computed, resulting in 11 features per frame.

The statistical properties of the pitch period are used in the LCQA algorithm and pitch estimation in low SNR environments is a challenging task, where current algorithms may fail to perform reliably in such conditions [10]. For the purpose of intelligibility estimation in very noisy speech, pitch information obtained through the YIN algorithm [11] was found to correlate poorly with the subjective score. Given the computational complexity of the pitch tracker, and the poor robustness of pitch estimation algorithms in noisy speech, pitch has not been included as a feature.

3.1.2 Global Features

The per frame features are transformed into N per utterance features by modeling the statistical properties of the per frame features through the mean, variance, skewness and kurtosis operators. This statistical description gives a global description of the per frame features and helps to reduce considerably the dimension of the feature set.

3.1.3 Dimensionality Reduction

In order to improve the performance of the classification, it is necessary to retain those features that model the various properties of the signal most effectively. We apply a two step dimensionality reduction scheme based on a feature subset selection followed by a feature extraction step on the training
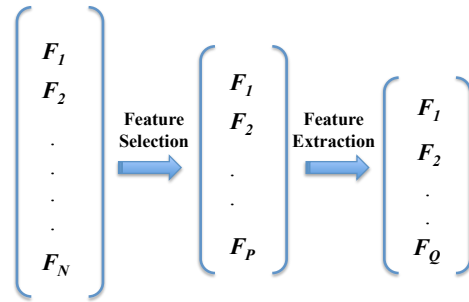

Figure 1: The dimensionality reduction scheme involves a feature selection (correlation) followed by feature extraction (PCA).

data, as shown in Fig. 1. The first stage is a feature subset selection, which is achieved through a correlation analysis of the features. It is desirable to retain only those features that have a high correlation with the intelligibility and at the same time, are uncorrelated with other features. The correlation coefficient based measure for feature i is obtained as :

$$Cor_i = \frac{R_i}{\sum_{j=1}^{N} R_{ij}}, \qquad (1)$$

where N is the number of features in the global set before feature selection and R_i is the correlation of the feature i with intelligibility scores and R_{ij} is the correlation of feature i with feature j. The correlation coefficient between vectors $\hat{I}$ and I is defined as:

$$R = \frac{\sum_n (\hat{I}_n - \mu_{\hat{I}})(I_n - \mu_I)}{\sqrt{\sum_n (\hat{I}_n - \mu_{\hat{I}})^2 \sum_n (I_n - \mu_I)^2}}, \qquad (2)$$

where μ_I and $\mu_{\hat{I}}$ denote the mean of I and $\hat{I}$ respectively. The correlation coefficient based measure is optimized to select P features with the highest correlation coefficient Cor_i from the set of N global features. The second step is a feature extraction, where PCA is used to transform the P features into Q dimensions by a linear combination ($N > P > Q$). In our experiments described later in this paper we have shown examples for the illustrative case of $P = 8$ and $Q = 7$.

3.1.4 Gaussian Mixture Modeling

A joint GMM is trained on the Q extracted features and the intelligibility score for each speech utterance in the training data. The GMM was tested with a range of mixtures and the optimal number of mixtures was found to be 7, giving the highest correlation and lowest MSE of estimated intelligibility with subjective scores (determined experimentally).

3.2 Importance weighted signal-to-noise ratio (*iSNR*) feature

The signal-to-noise ratio is a popular objective measure for quantifying the amount of additive noise in the signal. We use an intelligibility specific frequency weighted SNR measure to quantify effects of additive noise for each time frame of the signal. This forms a per frame feature whose statistical properties over the entire utterance is evaluated. The
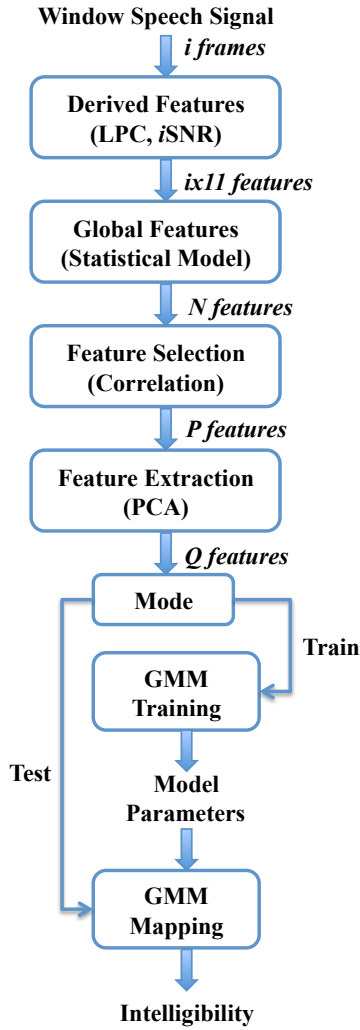
Figure 2: Illustration of the modified LCQA algorithm optimized for intelligibility estimation.

noise power is estimated using the minimum statistics algorithm [12] for each frame of the signal. The algorithm assumed an additive noise model:

$$x(n) = s(n) + v(n), \quad (3)$$

where $x(n)$ is the noisy speech, $s(n)$ is the speech signal and $v(n)$ is the noise.

The SII [1] is an intrusive measure that quantifies the aspects of the signal that are audible and usable to the listener. The SII score is monotonically related to intelligibility and is given in the range 0 to 1. The SII describes different Frequency Importance Functions (FIFs) based on different speech material. The FIFs are weighting functions applied to the signal spectrum based on the importance of the particular frequency band to intelligibility. The general SII formula is defined as:
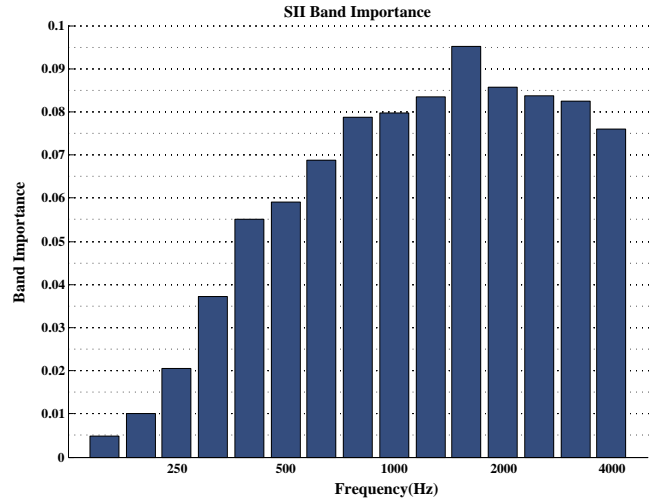
$$\text{SII} = \sum_{k=1}^{N_f} I(k)A(k), \quad (4)$$

Figure 3: 1/3rd octave band importance function used in the SII calculation [1].

where, N_f is the number of frequency bands. The band importance function $I(k)$, describes the importance of a frequency band to speech intelligibility and $A(k)$ is the band audibility function [1].

The $iSNR$ for frame i is defined as:

$$iSNR(i) = 10 \sum_{k=1}^{N_f} I(k) \log_{10} \frac{\max(0, P_x(i,k) - P_{\bar{v}}(i,k))}{P_{\bar{v}}(i,k)}, \quad (5)$$

where $P_x(i,k)$ is the power spectrum of the input (noisy speech) signal, computed as follows:

$$P_x(i,k) = X(i,k)X^*(i,k), \quad (6)$$

where $X(i,k)$ is the Discrete Fourier Transform (DFT) of the ith frame of the input signal. The estimated noise power $P_{\bar{v}}(i,k)$ is obtained in a similar way. It is important to estimate the $iSNR$ only for those periods in which the speech signal is active. The $iSNR$ calculation is thus restricted to voiced frames of the signal.

3.3 Database

The database consists of 200 sentences [13] from a male speaker. The sentences were corrupted with dynamic samples of car and babble noise at five SNRs obtained to correspond to an SII score of 0.1, 0.3, 0.5, 0.7 and 0.9. The speech activity level was obtained through the ITU-T P.56 algorithm [14] and this was used in the SNR calculation when adding the noise. Also included in the database are the noisy utterances processed through the spectral subtraction algorithm [15, 12] available in the Voicebox toolbox [16]. The 20 conditions in the database are summarized in Table 1.

Subjective intelligibility results were obtained from 20 naïve native speakers of British English. All subjects had hearing thresholds of less than 20 dBHL at frequencies ranging from 125 Hz to 8 kHz. The task was to listen to the stimuli and give a vocal reply which was recorded and scored. There were 5 keywords per sentence for the subject to identify. The subjective scores were averaged over the conditions

Condition	Noise	SNR (dB)	Suppression
1	Car	-9	off
2	Car	-12	off
3	Car	-15	off
4	Car	-18	off
5	Car	-21	off
6	Babble	0	off
7	Babble	-3	off
8	Babble	-6	off
9	Babble	-9	off
10	Babble	-12	off
11	Car	-9	on
12	Car	-12	on
13	Car	-15	on
14	Car	-18	on
15	Car	-21	on
16	Babble	0	on
17	Babble	-3	on
18	Babble	-6	on
19	Babble	-9	on
20	Babble	-12	on

Table 1: Database conditions, the suppression refers to processing the noisy speech with the spectral subtraction algorithm.

to give a condition averaged word intelligibility score in the range 0 to 1. As the same speaker was used for all the utterances, speaker independence has not been investigated in the current study.

3.3.1 Training

The database was partitioned into a test set and a training set. The speech material used in the training set was not included in the test set. Two training schemes were employed:

- 50% cross validation – in this scheme, we partition the database into an equal dimension test and training set. The training set contains all the conditions that are present in the test set. However, the test speech material is not available in the training set. The test and training set are swapped and the performed is the average over the cross validated sets.

- Predicting processing effects – in this scheme, the training set only contains the noisy speech conditions and has no example of the speech processed through spectral subtraction. Here we are interested in investigating the ability of the algorithm to predict the effects of speech enhancement on intelligibility.

4. RESULTS

We describe two experiments based on the training schemes presented in the previous section. For the purpose of these experiments, it has been found that selecting 8 features from the 40 global features and 7 linear combinations after the feature extraction give good results ($N = 40$, $P = 8$ and $Q = 7$). A non-linear relationship is known to exist between percentage correct intelligibility scores and SII [1]. Therefore a performance metric that accounts for this must be used. The Spearman rank correlation coefficient [17] is

Global Feature	Correlation
Skewness(spectral dynamics)	0.90
Kurtosis(spectral dynamics)	0.86
Skewness(d/dt(excitation variance))	0.80
Skewness(d/dt(iSNR))	0.61
Skewness(excitation variance)	0.59
Kurtosis(d/dt(excitation variance))	0.59
Skewness(d/dt(spectral centroid))	0.57
Kurtosis(iSNR)	0.57

Table 2: Table showing the absolute correlation coefficients for the raw features with subjective intelligibility scores (computed individually).

	Subjective	SII	LCIA
Subjective	1.0		
SII	0.91	1.0	
LCIA	0.92	0.96	1.0

Table 3: Correlations for the 50% cross validation partitions (all test conditions are present in training).

a non-parametric measure that describes the monotonic relationship between two variables, unlike the Pearson correlation coefficient (2) which describes a linear relationship. The Spearman correlation coefficient (ρ) is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \qquad (7)$$

where d_i is the difference between the statistical rank of the subjective and estimated intelligibility scores. The performance of the SII is compared with our non-intrusive intelligibility method, LCIA that is based on the LCQA algorithm.

4.1 Training on all conditions

In the 50% cross validation training scheme, examples of all the 20 conditions are represented in the training and test sets. The results from this experiment are presented in Table 3. The LCIA results have a correlation of 0.96 with the ANSI standard SII algorithm. This confirms that the modeling within LCIA has a well defined behavior. Also, with 0.92 correlation with subjective intelligibility scores, the algorithm outperforms the SII in estimating the intelligibility for additive noise and spectral subtraction, even though LCIA is non-intrusive.

Also, the statistical properties of the spectral dynamics is the most important feature (with a correlation of 0.90 with intelligibility) suggesting that the rate of change of the spectrum provides important information in intelligibility estimation.

4.2 Predicting processing

In this experiment, the training set only contains examples of the noisy speech and no examples of the speech enhanced through spectral subtraction. The algorithms are evaluated for their capability in predicting the effect of spectral subtraction on intelligibility. The results are shown in Table 4. For this scenario, the SII algorithm performs best, with a correlation of 1.0 with subject scores. The LCIA algorithm also has a high correlation of 0.96 with subjective scores.

	Subjective	SII	LCIA
Subjective	1.0		
SII	1.0	1.0	
LCIA	0.96	0.96	1.0

Table 4: Correlations with different test/train partitions (predicting effect of algorithm).

5. CONCLUSIONS

A low complexity data driven, non-intrusive speech intelligibility estimation algorithm was presented. The algorithm computes 40 features per utterance and applies a two step dimensionality reduction based on correlation and PCA. This results in 7 features, which are used to train a GMM of 7 mixtures. The statistical modeling of the features through skewness and kurtosis were found to correlate well for speech corrupted by noise and for predicting the effects of spectral subtraction. Also, the importance function weighted signal-to-noise ratio was presented as an important feature.

The algorithm has a correlation of 0.96 with the intrusive SII method and it was shown to predict the effects of processing after spectral subtraction with a correlation of 0.96. Finally, our approach was shown to give a correlation of 0.92 with subjective intelligibility scores.

REFERENCES

[1] ANSI, "Methods for the Calculation of the Speech Intelligibility Index," American National Standards Institute, ANSI Standard S3.5-1997 (R2007), 1997.

[2] ——, "Methods for the Calculation of the Articulation Index," American National Standards Institute, New York, ANSI Standard ANSI S3.5-1969, 1969.

[3] V. Grancharov, D. Zhao, J. Lindblom, and W. Kleijn, "Low-Complexity, Nonintrusive Speech Quality Assessment," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1948–1956, 2006.

[4] ITU-T, "ITU-T coded-speech database," ITU-T Supplement P.Sup23, Feb. 1998.

[5] ——, "Single-ended method for objective speech quality assessment in narrow-band telphony applications," ITU-T Recommendation P.563, 2004.

[6] S. Stearns, "On selecting features for pattern classifiers," in *Proc. 3rd Int. Conf. Pattern Recognition*, 1976, pp. 71–75.

[7] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. IEEE Int. Conf. Pattern Recognition*, 1994, pp. 279–283.

[8] ITU-T, "Methods for subjective determination of transmission quality," Online, ITU-T Recommendation P.800, Aug. 1996. [Online]. Available: http://www.itu.int/rec/T-REC-P.800/en

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[10] D. Sharma and P. A. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in *Proc European Signal Processing Conf*, 2009.

[11] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans on Speech and Audio Processing*, vol. 9, pp. 504–512, Jul. 2001.

[13] M. W. Smith and A. Faulkner, "Perceptual adaptation by normally hearing listeners to a simulated hole in hearing," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4019–4030, 2006.

[14] ITU-T, "Objective Measurement of Active Speech Level," ITU-T Recommendation P.56, Mar. 1993.

[15] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," vol. 4, 1979, pp. 208–211.

[16] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[17] E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*. Englewood Cliffs, NJ: Prentice-Hall, 1998.