NON-INTRUSIVE SPEECH INTELLIGIBILITY ASSESSMENT

Dushyant Sharma[1†] *, Patrick A. Naylor*[2] *and Mike Brookes*[2]

[1] Nuance Communication Inc,
Marlow, UK
email: dushyant.sharma@nuance.com

[2] Imperial College,
London, UK
email: p.naylor@imperial.ac.uk

ABSTRACT

We present NISI, a novel non-intrusive speech intelligibility assessment method based on feature extraction and a binary tree regression model. A training method using the intrusive STOI method to automatically label large quantities of speech data is presented and utilized. Our method is shown to predict speech intelligibility with an RMS error of 0.08 STOI on a test database of noisy speech.

***Index Terms*—** Speech Intelligibility, Speech Quality, Classification and Regression Trees, Data Driven

1. INTRODUCTION

The performance of many speech processing systems worsens in the presence of signal degradations. In law enforcement audio collection, very severe degradations can arise and reduce the intelligence value of audio by making it unintelligible or inadmissible in a court of law [1].

The perceptual effects of distortions on the speech signal are typically measured through speech quality assessment techniques [2]. Certain speech assessment techniques consider speech intelligibility to be an aspect of speech quality, as in the diagnostic acceptability measure [3]. It is an important quantifier for applications such as telecommunications, where a channel may be evaluated in terms of its effect on speech intelligibility [4], as a performance metric for hearing aids [5], for determining the impact of an acoustic space on speech [6] and for intelligence gathering in law enforcement applications [7].

Speech intelligibility can be defined as a measure of the proportion of a speech signal's content correctly recognised by a listener. A number of methods have been proposed in the literature for obtaining speech intelligibility scores and these may be classified as either subject-based or objective measures. Subject-based speech intelligibility scores are obtained through listening experiments where subjects listen to speech samples and their performance in a particular linguistic task is measured. The linguistic task may be to recognize nonsense syllables, isolated words or specific keywords in a sentence.

Objective intelligibility assessment methods operate without the need for human listeners and can provide rapid intelligibility scores. These can be further divided into two classes: (i) those requiring a reference signal in addition to the test signal are referred to as intrusive methods, and (ii) those operating only on the signal under test are referred to as non-intrusive methods.

One of the earliest intrusive intelligibility technique was proposed by French and Steinberg [8] as the Articulation Index (AI), which was further developed into the Speech Intelligibility Index (SII) and led to an ANSI standard [9]. The SII evaluates the effects of degradations in a number of frequency bands, weighted by their importance to speech intelligibility, and quantifies the proportion of the speech signal that is intelligible to the listener. The SII score is monotonically related to the intelligibility of the speech utterance and is given in the range 0 to 1 (where a score of 0.5 means that half of the speech cues are audible and usable to the listener) [10]. More recently, the Short-Time Objective Intelligibility (STOI) method for intrusive intelligibility assessment has been proposed which has been shown to have a high correlation (better than 0.92) with subjective intelligibility scores for both noisy and noise-suppressed speech [11].

The low complexity speech intelligibility method (LCIA) [12] is a data-driven non-intrusive measure that has been shown to have a high per-condition correlation with subjective intelligibility scores of noisy and noise suppressed speech. The LCIA method has been evaluated using subjective sentence intelligibility scores from [13] but this provided only a limited number of degradation conditions.

In this paper we propose the non-intrusive speech intelligibility (NISI) method and evaluate its performance on a large database of noisy speech labeled with intelligibility scores using the intrusive STOI [11] method. The use of an intrusive method such as STOI in this case to label the database automatically instead of subject-based intelligibility labelling has the advantage of allowing for a large training and evaluation of non-intrusive methods to be performed at substantially lower financial and time costs.

The remainder of this paper if organized as follows. Section 2 reviews the LCIA method and the NISI method is

[†] Dushyant Sharma was a PhD student at Imperial College London during the course of this work.

presented in Section 3, followed by evaluation methodology, databases and metrics in Section 4. The results are presented in Section 5 followed by conclusions in Section 6.

2. LCIA REVIEW

The LCIA method [12] is a data-driven approach for low-complexity, non-intrusive speech intelligibly assessment; it is a development of the LCQA method [14] with a new feature, importance weighted SNR (iSNR), an external Voice Activity Detector (VAD), the use of a two-step feature selection and projection technique and is trained on speech data labeled with intelligibility scores. The LCIA method begins by deriving per frame features from the speech waveform, then applying a statistical model followed by a two-step dimensionality reduction and Gaussian Mixture Model (GMM) mapping. In contrast to LCQA, the pitch period is not used as a feature in LCIA due to the computational complexity of pitch tracking, and the poor correlation of this feature with subjective intelligibility scores [12], particularly for highly degraded audio. The features are listed in Table 1. The statistics of the per-frame features results in a 44 dimensional feature vector per utterance, which is further reduced in dimension by a correlation based feature selection and principal component analysis (PCA) based feature projection. The intelligibility score is obtained from the output of a joint GMM (diagonal covariance matrix), trained on the projected features and the intelligibility score for each speech utterance in the training data.

3. NISI METHOD

The Non-Intrusive Speech Intelligibility (NISI) method is a data-driven, machine learning approach to speech intelligibility estimation whose overall structure is presented in Fig. 1. The first step is a short-time segmentation of the input signal into 20 ms frames employing a non-overlapping Hanning window, denoted $y(i)$, where i is the frame index. This is followed by VAD based on the P.56 method [15] to select frames where speech is present. This is then followed by short-term feature extraction and the statistics of the short-term features (mean, variance, skewness and kurtosis) are used to characterize the entire signal and combined with the long-term features to create the final feature vector of dimension 116. The features form the input to a CART regression model that has been previously trained on a feature matrix with corresponding ground truth scores. The features used in NISI method are listed in Table 1 and further described in the following subsections.

3.1. Short-term features

The short-term feature extraction follows the time segmentation of the input speech signal into voice-active frames and are described as follows. The NISI method makes use of pitch estimates, and rate of change of pitch, obtained from the new PEFAC algorithm [16] that has been shown to be robust to additive noise. The iSNR [12] feature is an intelligibility specific frequency weighted measure to quantify the effects of additive noise in the signal and is a feature in NISI along with the rate of change of iSNR over the utterance. A 10th order linear predictive coding (LPC) is performed on the speech signal, and the residual variance and its rate of change over the utterance are included as features. Additionally, the spectral centroid, flatness and dynamics of the LPC magnitude spectrum are computed, as in [14], and included in the feature vector along with their rates of change. The zero crossing rate and its rate of change over the utterance are also used as features. The NISI method utilises features based on the Hilbert envelope:

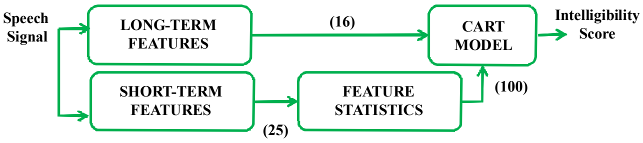

Fig. 1. Overview of the NISI method. The dimension of the features after each block is shown in ().

$$e(i) = \sqrt{y(i)^2 + \mathscr{H}(y(i))^2}, \tag{1}$$

where $e(i)$ is the envelope of the i^{th} frame of $y(n)$ and $\mathscr{H}\{.\}$ is the Hilbert transform. The variance ($\sigma_{e(i)}$) and dynamic range ($\Delta_{e(i)}$) of the envelope for each of the N_i frames are computed as follows and used as features:

$$\sigma_{e(i)} = \frac{1}{N_i}\sum_{i=1}^{N_i}(e(i) - \mu_{e(i)})^2 \tag{2}$$

$$\Delta_{e(i)} = |\max(e(i)) - \min(e(i))|. \tag{3}$$

Additionally, the rates of change of these features are included.

The long term average speech magnitude spectrum (LTASS) is often used as a model for the clean speech spectrum. The power spectrum of long term deviation (PLD) feature for frame i and frequency bin k is defined as:

$$\text{PLD}(i,k) = \log(P_y(i,k)) - \log(P_{LTASS}(k)), \tag{4}$$

where $P_y(i,k)$ is the magnitude power spectrum of noisy signal and $P_{LTASS}(k)$ is the LTASS power spectrum.The per-frame PLD spectrum is used to derive the spectral flatness (PLD Flatness), spectral centroid (PLD Centroid) and spectral dynamics (PLD Dynamics) features. The spectral flatness, dynamics and centroid of PLD spectrum and their rate of change are included as short-term features in NISI.

	LCIA		NISI	
	ϕ	$\Delta\phi$	ϕ	$\Delta\phi$
LPC Flatness	ϕ_1	ϕ_7	ϕ_1	ϕ_{14}
LPC Dynamics	ϕ_2		ϕ_2	
LPC Centroid	ϕ_3	ϕ_8	ϕ_3	ϕ_{15}
LPC Residual	ϕ_4	ϕ_9	ϕ_4	ϕ_{16}
Speech variance	ϕ_5	ϕ_{10}	ϕ_5	ϕ_{17}
Pitch period			ϕ_6	ϕ_{18}
Zero crossing rate			ϕ_7	ϕ_{19}
iSNR	ϕ_6	ϕ_{11}	ϕ_8	ϕ_{20}
Envelope variance			ϕ_9	ϕ_{21}
Envelope range			ϕ_{10}	ϕ_{22}
PLD Flatness			ϕ_{11}	ϕ_{23}
PLD Dynamics			ϕ_{12}	ϕ_{24}
PLD Centroid			ϕ_{13}	ϕ_{25}
$\mathrm{P_{LTLD}}$			$\phi_{26:41}$	

Table 1. The features used in the LCIA method ($\phi_{1:11}$) and NISI method ($\phi_{1:41}$). The columns labelled ϕ and $\Delta\phi$ denote the raw features and their time derivatives respectively.

3.2. Long-term features

The long-term deviation of the magnitude spectrum of the signal (calculated over the entire utterance) is defined as follows

$$\mathrm{P}_{LTLD}(k) = \frac{1}{N_i}\sum_{i=1}^{N_i}\mathrm{PLD}(i,k). \qquad (5)$$

The resulting P_{LTLD} spectrum is mapped into 16 bins each with a bandwidth of 500 Hz and 50 % overlap. The energy in each bin as a percentage of the total energy is then computed to form the long term features in NISI. It is intended that this feature can identify the long-term frequency characteristics of different types of degradations.

3.3. Classification and Regression Tree

The CART algorithm [17] is used to construct a regression tree for modeling speech intelligibility using the previous feature extraction mechanism. CART recursively partitions the feature space using binary splits into a number of terminal nodes, each containing a constant predicted response value. In this use of CART, first, an over-sized and sub-optimal tree is grown using a minimum mean square error (MSE) criterion. This is followed by a pruning process using 10-fold cross-validation of the training data to merge tree branches that result in small reduction in the MSE.

4. EVALUATION

4.1. Databases

The training and validation database (referred to here as the TN database) is based on the TIMIT database [18], which contains speech from 623 speakers of US English. In the TN database, only the distinct utterances are used for all the speakers. An extensive additive noise database is then created by adding 15 noises from the NATO noise database [19] at SNRs in the range -24 to 30 dB in 3 dB steps. The resulting TN database is composed of 285 additive noise conditions for each speaker.

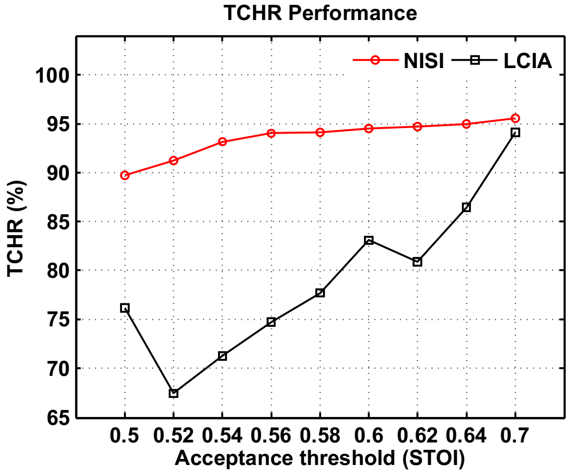

Fig. 2. STOI estimation performance using the TCHR metric.

The additive noise partition of the C-Qual database [20], comprising of car, babble and hum noise representing 21 conditions for each of 4 speakers is used only as a generalization test database as it contains different speech and noise material from the TN database. All databases were down-sampled to 8 kHz to represent narrow-band speech transmission.

4.2. Training

The original TIMIT database is partitioned into a training and test partition; this is maintained in the TN database, with the training partition consisting of 168 speakers randomly selected from the original 455 speakers in the TIMIT training partition. All data-driven algorithms are trained on the TN training partition and also tested on the entire test partition. Additionally, each noise file was also split into a training and test partition so that the noise source samples in the training and test partitions are different. The resulting TN database consists of more than 46 hours of speech material.

4.3. Evaluation Metrics

4.3.1. Spearman Correlation Coefficient (SCC)

The Spearman rank correlation coefficient (SCC) is a nonparametric measure that describes the monotonic relationship between two ranked variables [21] in the range -1 to +1.

4.3.2. *Root Mean Square Error (RMSE)*

The root mean square error between the estimated and true scores is calculated as a measure of the estimation accuracy of each algorithm.

4.3.3. *Bin Error*

This measure evaluates the absolute mean residual error in the true and estimated STOI scores in bins of size 0.05 STOI, by dividing the STOI scale into 20 bins. This metric shows the percentage of signals that lie in each bin and provides a histogram view of the errors.

4.3.4. *Two Class Hit Rate (TCHR)*

This measure investigates the hit rate achieved by splitting the ground truth scores into two classes (according to an acceptance threshold). The acceptance threshold is set to the STOI score corresponding to 75% intelligibility, as provided by the mapping function proposed in [11] to be 0.62 STOI. The motivation for this comes from a previous study [22] where a threshold of acceptance at 75% was found to be practical for intelligibility assessment. The TCHR metric is also evaluated at a number of other threshold values to assess how the performance changes with different threshold values. Thresholds in the 0.5 to 0.7 STOI range are evaluated, corresponding to word intelligibility scores from 28% to 93% words correct.

5. RESULTS

5.1. Performance for TN database

Table 2 shows the performance of the LCIA and NISI methods in estimating STOI on the TN database. The NISI method outperforms LCIA on all metrics tested, achieving an SCC of 0.95 and an RMSE of 0.08 STOI. The NISI method has a high accuracy, with 93.3% of errors less than 0.15 STOI and for an acceptance threshold of 0.62 STOI, the TCHR performance is nearly 95%. The LCIA method also has a high correlation in this task (SCC = 0.91) but a poor estimation accuracy, with an RMSE of 0.18. The performance of the methods for different acceptance thresholds is presented in Fig. 2, where NISI can be seen to have a consistent performance with a TCHR higher than 90% in the region of 0.5 to 0.7 STOI (28% to 93% intelligibility). The 5 best ranked features for STOI estimation are presented in Table 3, where the iSNR, LPC dynamics, PLD Flatness are seen to be most important short-term features. Also, the long-term PLD based feature, ϕ_{27} (deviation in the 250-750 Hz band) is important.

5.2. Performance for C-Qual database

The generalization performance of the methods is presented in Table 4, where the methods are trained on the TN database and tested on the additive noise partition of the C-Qual database. The overall performance for both methods in this task is lower, with the best performance provided by NISI (SCC of 0.86 and RMSE of 0.12). This may be partly due to the differences in the types of degradations between the C-Qual and TN databases.

			Bin Error (%)			TCHR
	SCC	RMSE	< 0.1	< 0.15	< 0.2	@0.62
NISI	**0.95**	**0.08**	**85.4**	**93.3**	**97.0**	**94.7**
LCIA	0.91	0.18	45.4	61.3	72.9	80.9

Table 2. STOI estimation on the TN database.

Rank	LCIA	NISI
1	$\mu(\phi_5)$	$\mu(\phi_8)$
2	$\sigma(\phi_6)$	$\sigma(\phi_2)$
3	$\mu(\phi_6)$	$\sigma(\phi_{23})$
4	$\sigma(\phi_2)$	$\mu(\phi_3)$
5	$\mu(\phi_2)$	ϕ_{27}

Table 3. 5 best ranked features for TN database. The mean (μ) and variance (σ) are important statistics.

			Bin Error (%)			TCHR
	SCC	RMSE	< 0.1	< 0.15	< 0.2	@0.62
NISI	**0.86**	**0.12**	**70.2**	**84.5**	**90.5**	**88.1**
LCIA	0.82	0.15	46.4	67.9	84.9	79.8

Table 4. STOI estimation on the C-Qual database.

6. CONCLUSIONS

The non-intrusive assessment of speech intelligibility was considered in this paper. A novel data-driven method, NISI, was presented and shown to correlate strongly with STOI. Moreover, the performance of NISI was shown to be highly consistent in the two class classification task, achieving a hit rate higher than 90% over a large range of intelligibility scores. The proposed PLD based features were shown to be important for predicting the STOI scores in the TN database. The LCIA method was further evaluated for predicting STOI scores of noisy speech. A novel technique for automatically labeling databases for speech intelligibility using the intrusive STOI method was also presented. The TN database was developed by adding 15 noises at SNRs in the -24 to 30 dB range to clean speech from the TIMIT database. The resulting database was split into a test and training partition with no overlap of noise source, speakers or speech material and the LCIA and NISI methods were evaluated on this database.

7. REFERENCES

[1] R. C. Maher, "Audio forensic examination," *IEEE Signal Processing Magazine*, vol. 26, pp. 84–94, 2009.

[2] S. Moller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, pp. 18–28, 2011.

[3] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1977, pp. 204–207.

[4] 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, Jan. 1980.

[5] R. J. M. van Hoesel and R. S. Tyler, "Speech perception, localization, and lateralization with bilateral cochlear implants," *J. Acoust. Soc. Am.*, vol. 113, no. 3, pp. 1617–1630, 2003.

[6] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[7] P. A. Naylor, N. D. Gaubitch, D. Sharma, G. Hilkhuysen, M. Huckvale, and M. Brookes, "Intelligibility estimation in law enforcement speech processing," in *Proc ITG Conf on Speech Communication*, Bochum, Germany, Oct. 2010.

[8] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[9] ANSI, "Methods for the calculation of the speech intelligibility index," American National Standards Institute, ANSI Standard S3.5–1997 (R2007), 1997.

[10] B. W. Y. Hornsby, "The Speech Intelligibility Index: What is it and what's it good for?" *Hearing Journal*, vol. 57, no. 10, pp. 10–17, October 2004.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, September 2011.

[12] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Denmark, Aug. 2010.

[13] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility: dependency on signal-to-noise ratios," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 531–539, 2012.

[14] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

[15] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[16] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Aug. 2011.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. CRC Press, 1984.

[18] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[19] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 3, no. 3, pp. 247–251, Jul. 1993.

[20] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, M. Brookes, and P. A. Naylor, "C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio," in *Proc. AES Conf. on Audio Forensics*, Hillerød, Denmark, Jun. 2010.

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

[22] K. Worrall and R. Fellows, "Practical and affordable intelligibility testing for engineers and algorithm developers," in *Proc. AES Conf. on Audio Forensics*, Hillerod, Denmark, June 2010, pp. 194–201.