Proceedings of the Institute of Acoustics

THE COMPLEXITY OF SPEECH INTELLIGIBILITY MEASUREMENTS IN PACKETISED TRANSMISSION CHANNELS

Francis F. Li, Manchester Metropolitan University

1. INTRODUCTION

The integration of computer and telecommunications technologies has enabled the transmission of multimedia signals over packetised data communications networks locally and globally. With the rapid growth of the Internet, Voice over Internet Protocol (VoIP) technology has become a potential alternative to and supplement of the traditional Public Switched Telephone Network (PSTN), offering a versatile, flexible and cost-effective solution to peer-to-peer speech communications. VoIP integrates voice, multimedia and data communications in a unified framework and infrastructure and hence provides, hypothetically, a manageable and scalable network infrastructure to satisfy the needs of various communications. It can also be used in conjunction with PSTN landlines and wireless communications networks to provide an affordable solution, under the current billing regime, to long distance calls.

Quality of Service (QoS) is an important concern of any Internet based services. The ultimate goal of a telephony system, regardless of its transmission channels or media, is to deliver intelligible speech communications. Speech quality, particularly speech intelligibility, naturally becomes a key issue of various services around VoIP. The lack of QoS guarantee hinders the widespread applications of VoIP. An effective and objective assessment method is sought as an essential step towards assured quality. The significance is threefold: first, it offers a diagnostic means to identify problems of transmission channels; secondly, a quantified specification can be included in the service agreements and finally an intelligibility monitor may be used to reassure users that the QoS is achieved.

Quantifying the quality of a speech transmission channel in terms of its intelligibility is by no means easy. Time-variance and the discontinuity of a packetised network cause additional complexity in VoIP channels. Low-bit-rate non-linear codecs at front ends of a VoIP channel further aggravate the complexity. A review has revealed the limitations of existing methods in VoIP applications: due to the time variance, subjective methods become unrealistically time-consuming and complicated; existing objective methods do not effectively evaluate echo and packet loss impacts on speech. Moreover, objective methods are preferable to

subjective ones in the VoIP context in general for their good repeatability, fast evaluation and capability to perform long-term monitoring.

This paper will discuss the complexity of extending existing speech intelligibility assessment methods to VoIP applications and propose a framework for machine assessments of speech intelligibility of VoIP channels.

2. OBJECTIVE VERSUS SUBJECTIVE MEASUREMENTS

Speech intelligibility, in a broad sense is a subjective judgement indicating how well the information carried by a speech signal can be 'decoded' by a human listener. Subjective tests play an irreplaceable role in speech intelligibility assessments. However, articulations of speech, characteristics of transmission channels, noise levels, acoustic conditions all have implications on intelligibility of transmitted speech. Moreover, how a listener perceives a particular speech can also affect the subjective judgement. Does a transmission channel have identical subjective intelligibility for different languages? Will a channel show identical intelligibility to users of all age groups? These make individual subjective test results less meaningful and hence less useful as a specification for QoS. Statistical approaches should be adopted using a significant number of trained talkers and representative listeners. Moreover, multi-language tests are needed to fully assess the intelligibility of a system. These tests are known to be extremely time-consuming and complicated to set up and carry out. How to perform statistical analysis and interpret the results is an additional challenge.

VoIP is based on packetised networks. Network traffic and congestion are beyond the control of system operators and users, resulting in severe time variance of VoIP channels. Different codecs used at user ends complicate the scenario. Individual test results can be unrepresentative and are often unreliable. Speech intelligibility for VoIP needs long-term monitoring to give a statistically meaningful profile, which makes subjective tests even more time-consuming and difficult to carry out.

Speech intelligibility of a transmission channel, which is the impact of transmission channels on transmitted speech, is of QoS interest. It is a description of certain physical characteristics of the channel rather than the speech. Such a description should have a strong correlation with subjective perception of transmitted speech.

3. EXISTING METHODS

Traditionally two different types of assessment methods are used, subjective and objective. Subjective intelligibility assessments are based on scores for correct recognition of utterances from pronounced phonemes, words, word lists, sentences or quality rating according to general subjective impression of listeners. In contrast, objective measurements use specific test signals and equipment to perform

measurements. The significant attractiveness of objective methods is that they are on physical grounds and may be related to physical characteristics of speech transmission channels.

Due to the severe time variance of VoIP systems, full subjective intelligibility tests are far too time-consuming and complicated to carry out. Subjective quality rating is often used to give a coarse estimate. Mean Opinion Score (MOS) is probably the most popular subject method used for VoIP quality assessments. Transmitted sentences or a free conversation are used to obtain listeners' subjective impression. A five-grade scale, from 1 to 5, representing speech quality from very poor, poor, fair, good, to excellent, is used. A MOS score of higher than 3.5 indicates a good quality. The MOS method has three significant limitations in intelligibility assessment: (1) it reflects not only intelligibility but other auditory impressions as well, (2) the grades used are not calibrated and normalised, and (3) MOS scores often show large variations among different listeners.

A number of objective methods were previously developed to quantify speech intelligibility for rooms and other transmission channels such as PA and telephony systems. Articulation Index (AI) [1], Speech Transmission Index (STI) [2] and Speech Intelligibility Index (SII) [3] are well known in acoustic research community.

The AI method measures system responses in twenty crucial frequency bands and gives a weighted-score according to contributions of each frequency band to the speech intelligibility. Obviously, the AI method focuses on frequency distortions of transmission channels. It is useful in the assessment of a continuous transmission channel, where speech intelligibility is degraded predominately by frequency distortions. In a VoIP system, the IP network does not impose any frequency distortions on readily digitised and packetised speech signals. Although the AI method might be applicable to the assessment of certain kinds of codecs, it is not suitable for the assessment of end-to-end intelligibility of VoIP systems.

The STI method uses speech-like artificial test signals, focuses on envelope shaping effect at fourteen modulation frequencies in seven octave bands of speech interest and gives a single index showing a close correlation with subjective perception of intelligibility. The SII method is very similar to the STI method, but can take more frequency bands into account. It is known that STI and SII methods may become problematic in the presence significant late-arriving reflections and echoes. Significant echoes and packet-losses are known to be the major problems in VoIP transmissions. As a result STI and SII methods may not be applicable directly to VoIP channel assessments.

4. VOIP CHANNELS

Intelligibility of a speech transmission channel is related to channel transfer characteristics. In a linear channel, the impulse response and channel noise solely

determine the intelligibility. A number of methods can be used to measure the impulse responses. Intelligibility parameters such as STI can be derived from the impulse response under certain necessary conditions (linear, passive and time invariant). VoIP channels however are subject to packet-loss, transmission delays and echoes. Network performance has a significant impact on the quality of transmitted narrow band speech and is time-variant [4,5]. Although linear PCM codecs can be used, the widespread applications of discontinuous, low-bit-rate and non-linear codecs complicate the intelligibility assessment. For example, silence-suppression techniques such as the Voice Activity Detection (VAD) algorithm are used to prevent packets from being transmitted during the quiet periods between spoken phrases [6]. Various packet-loss protection mechanisms may be embedded in the codecs to solve the packet-loss problem [7,8,9], though the effectiveness is difficult to predict. The end-to-end VoIP channel is an extremely complicated chain. The time-variant and non-linear nature of such a chain stems from not only packetised network but also various intelligent algorithms used to achieve low-bit-rate transmission and packet-loss compensation. This makes it difficult to formulate an effective artificial test signal to probe the VoIP channel. Furthermore impulse responses are not applicable to such channels.

If the intelligibility of VoIP channels can be estimated from transmitted speech signals, many technical dilemmas can be resolved. A speech based approach can also satisfy the needs of long-time and in-use monitoring of QoS of VoIP.

Machine learning based methods to accurately determine STIs from received running speech signals were previously developed [9]. They were extended to the blind identification of STIs, i.e. estimation of STIs from received speech without monitoring the source [10]. It is postulated that these methods might be further extended to the intelligibility problem of VoIP.

5. PROPOSED SYSTEM

A pilot investigation has been carried out to study the applicability of the STI method to VoIP channels. Network congestion patterns are simulated using the well-known Gilbert model. Straightforward PCM codecs are used to avoid the complexity of codecs at the early investigation stage. Standard RaSTI method for telephony system as defined in the standards is experimented with. It is found that RaSTI data do not respond correctly to packet losses and echoes.

A framework derived from the machine learning based STI method as documented in [9 and 10], but with added echo and packet-loss detectors is proposed. Naturally occurring running speech, when a VoIP system is in use, is used as probe stimuli. The proposed method is illustrated as a block diagram in Figure 1. The proposed method examines the received speech envelopes. Hilbert Transform is applied to received speech signals to detect their envelopes. The envelopes are further

analysed to identify (1) the level of smearing and noises (2) echoes and (3) packet losses.

The quasi-RaSTI block as depicted in Figure 1 exploits the methods developed in [10]. It yields coarsely estimated RaSTI data form received speech signals. The echo detection block gives delay time and echo magnitude estimates. The packet-loss estimator monitors high frequency components in envelope spectra, since the packet-loss should cause abrupt cut-off of the signal flow and hence high frequency components in envelope spectra. (It is worth mentioning that the packet-loss estimator may not work with codecs using packet-loss compensation techniques.) The final stage performs weighting and normalisation. The weighting coefficients can be determined using a regression model. In the pilot study, a standard back-propagation neural network is adopted and trained on speech examples with know MOS scores. System output is normalised to the range of 1 to 5 representing 5 grades of the MOS scores. 160 examples were used in the training phase.

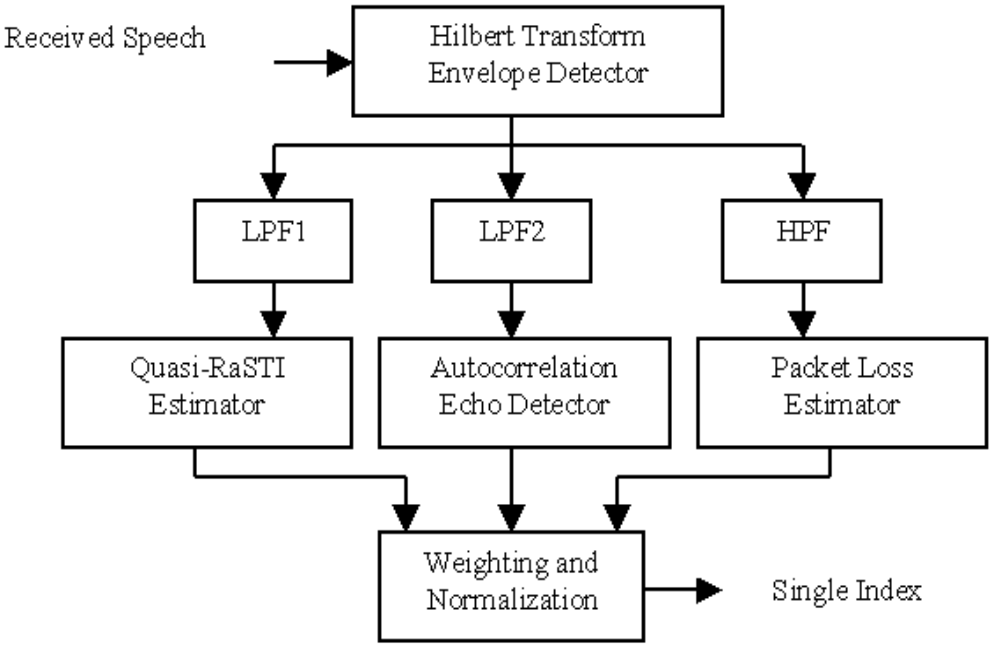

Figure 1. Proposed framework

4 sets of subjective tests were carried out, asking 4 listeners to give 1 to 5 graded subjective impression (MOS) of 20 speech examples. Although the numbers of training examples and tests are too small to give a statistically meaningful conclusion, the MOS scores obtained have shown high levels of correlation (correlation coefficient = 0.92) with the index obtained from the proposed system. The fining seems to suggest a possible pathway towards a new objective method for the assessment of intelligibility for VoIP.

6. CONCLUDING REMARKS

This paper has discussed the complexity of the measurement of intelligibility for VoIP channels. Traditional methods including AI, STI and SII are found inadequate for VoIP applications. Due to the time-variance and the need of long-time, in-use monitoring, a speech based method is proposed. A pilot investigation suggests that the proposed framework is feasible and worth further investigating.

7. ACKNOWLEDGEMENTS

The author would like to acknowledge the invitation from the Institute of Acoustics. Special thanks extend to the President of the IOA Mr. Geoff Kerry for his suggestions on the title of this paper.

8. REFERENCES

[1] K. D. Kryter. Methods for the calculation and use of the articulation index. *J. of the Acoustical Society of America*, 34:1689--1697, 1962

[2] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *J. Acoustical Society of America*, 67:318--326, 1980

[3] ANSI standard S3.5-1997

[4] L. A. R Yamamoto, J. G. Beerends, Impact of network performance parameters on the end-to-end perceived speech quality, Expert ATM Traffic Symposium, Greece, Sept. 1997

[5] N. S. Jayant and S.W. Christensen, Effects of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure, IEEE Trans. Commun., vol. 29: 101-109 no. 2, February 1981.

[6] ITU-T Recommendation. G.723.1, Annex A, "Silence compression scheme."

[7] N. Erdol et al., Recovery of missing speech packets using the short-time energy and zero-crossing measurements," IEEE Trans. Speech and Audio Processing, vol. 1, no. 3:295-303, July 1993.

[8] M. Yuito and N Matsuo, A new sample-interpolation method for recovering missing speech samples in packet voice communications, in Proc. ICASSP-89, pp. 381-384, 1989

[9] F. F. Li and T. J. Cox, “Speech transmission index from running speech: A neural network approach,” *Journal of Acoust. Soc. Am.*, Vol. 113, Issue 4, pp.1999-2008, 2003

[10] F. F. Li and T. J. Cox, “A neural network for blind identification of speech transmission index”, in the *Proceedings of IEEE ICASSP2003’*, v. II, pp. 757-760, 2003