INVESTIGATIONS INTO THE RELATIONSHIP BETWEEN MEASURABLE SPEECH QUALITY AND SPEECH RECOGNITION RATE FOR TELEPHONY SPEECH

Hanwu Sun, Louis Shue, Jianfeng Chen

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
Email: {hwsun,lshue,jfchen}@i2r.a-star.edu.sg

ABSTRACT

In this paper, an investigation to establish a possible relationship between the performance of a telephony speech recognition system and the method for objective speech quality assessment described in ITU-T Recommendation P.862, known as Perceptual Evaluation of Speech Quality (PESQ), is presented. Experiments using various additive background noises, and at different separations between the microphone and the sound-source have been conducted to establish such a relationship. The preliminary results suggest that telephony speech recognition rates can be mapped to the Mean-Opinion-Score (MOS) obtained by PESQ using a relatively simple polynomial relationship. This indicates that the PESQ MOS can act as a reliable predictor for the achievable speech recognition rates for telephony-based speech recognition system.

1. INTRODUCTION

Telephony-based speech recognition systems are gaining more popularity these days. Typical applications include automatic call centers, remote voice inquiries and commercial transactions. Currently, most commercial general-purpose speech recognizers can provide acceptable recognition accuracy for relatively clean speech. However, the performance of such recognition systems degrades significantly when they are applied to real situations where there are various noisy interferences [1–3]. Such problems may be more serious for the telephony-based recognition system, where the remote users may use speakerphone, or cellular/mobile phone in some noisy environments.

In 2002, the International Telecommunication Union (ITU) proposed a recommendation: ITU-T Recommendation P.862, also known as the Perceptual Evaluation of Speech Quality (PESQ) [4], for evaluation of quality of telephone-band (300-3400Hz) speech codecs. This recommendation made it possible to quantify channel errors that is more closely related to human perception. The ITU-T P.862 (see [5,6] for more in-depth discussions of the design considerations) replaces the previous ITU-T P.861 [7] and is able to provide a perceptually based quality evaluation for telephone-band speech in a range of conditions, including, for example, coding distortions, errors, noises, filtering, delay and variable delay. Subsequent experiments [1] comparing the PESQ Mean Opinion Score (MOS) and the generally used MOS listening tests [8] using human listeners have shown a strong correlation between the two methods.

The question we wish to address in the present research is: *is there a speech quality measure which can be used to predict speech recognition performance*? While it is generally acknowledged that signal-to-noise ratio (SNR) is not a good indicator of speech quality, see for example [9], in this paper, we will try to extend this notion further. In particular, we wish to use a measure applicable to speech recognition rather than speech enhancement, as in [9]. By making use of PESQ MOS, we firstly verify experimentally the relationship between recognition rates, SNR and speech quality for telephony speech. Next, by systematically introducing degradations in speech quality through

1. additive background noise, and
2. using data in a realistic office environment which will have undesirable convolutive effects,

so that a deterioration in speech recognition rates can be observed, we will establish a relationship between recognition rate and speech quality. The results suggest that the PESQ MOS can be used as a reliable predictor for the recognition performance[1] in a given situation. This seems to be consistent for both additive noises as well as any convolutive effects which exist in a realistic room.

The rest of the paper is organized as follows. The workings for the PESQ scheme is briefly reviewed in Section 2, followed by our proposed evaluation scheme in Section 3, in which we will also discuss the necessary tools and databases used. In Section 4, the experimental results will be discussed. Finally, some conclusions and suggestions for future work are given in Section 5.

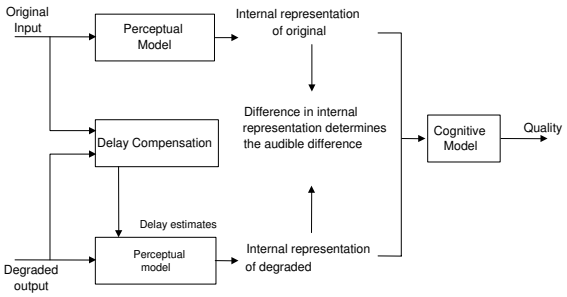


Fig. 1. ITU-T P.862 PESQ MOS evaluation procedure.

2. PESQ OVERVIEW

In this section, we will briefly describe the process of obtaining the PESQ MOS, more details can be found in [4–6].

[1] Or, more accurately, the likely decrease in recognition rate in the event of a decrease in speech quality.

The PESQ algorithm requires two input signals for a computation of speech quality: the original speech sample and a degraded version of it. These two signals are mapped into an internal representation using a perceptual model (see Fig. 1). The differences in the two representations are then used by a cognition model to predict the perceived speech quality of the degraded signal. The internal representations used by the PESQ cognition model to predict the perceived speech quality are calculated on the basis of signal representations that make use of the psychophysical equivalents of frequency (pitch measured in bark scale) and intensity (loudness measured in tones). Unlike the conventional MOS obtained from human listeners and uses scale of 1 to 5 (1=worst quality, to 5=best quality), the PESQ MOS ranges from -0.5 to 4.5 (-0.5=worst case and 4.5=best case, or no distortion). More significantly, it has been pointed out [6] that PESQ can be used to asses the quality of system carrying speech in presence of background or environmental noises (such as car or street noise).

3. EVALUATION FRAMEWORK

In this section, we will outline the experimental procedure and tools used in the subsequent experiments. A flowchart of our procedure is shown in Fig. 2.

3.1. Experimental Procedure

As shown in Fig. 2, G is a gain factor to adjust the additive noise in order to form a given SNR in our test speech signal. The SNR is defined as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{P_s - P'_n}{P_n}\right) \quad (1)$$

where P_s is the average power of segments of test signal containing speech, and P_n is the average power of the additive background noises respectively. In addition, P'_n is an estimate of the average noise power present during initial recording of audio samples for the speech database.

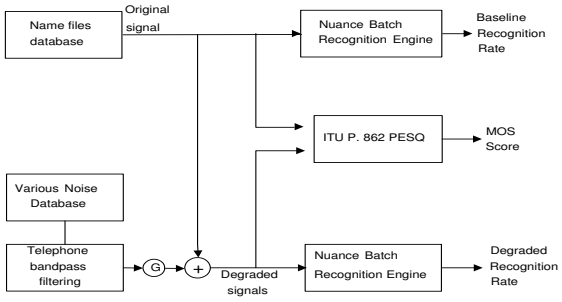


Fig. 2. Flowchart of the experimental procedure for establishing a relationship between PESQ MOS and speech recognition rates.

Two types of experiments were conducted to induce a reduction in speech quality. Firstly, additive noises at various SNRs were added and the resulting speech recognition performance monitored (see Section 4.1). Next, a degradation in speech quality was induced by placing microphones at increasing distances from the source, to evaluate the decrease in recognition rate due to convolutive effects. A high-end GENELEC 1029A loudspeaker was used for the playback of the audio samples in the speech database, which were then recorded simultaneously using four Knowles EA-2183 omni-directional microphones, placed at 10, 20, 40 and 80cm away from loudspeaker. These four microphone signals were amplified by using two TASCAM DA-P1 DAT recorders, each having two embedded microphone amplifiers. The babble noise was added to these audio samples digitally and the combined signals were used in the speech recognizer, see Section 4.2. We have made the assumption that the babble noise is in the far-field relative to the distances between the loudspeaker and four microphones.

3.2. The Tools

Speech Recognition Engine:

The Nuance 8.0 speech recognition engine (in batch mode and tuned specifically for telephony recognition) was used as the basic platform for evaluating the speech recognition performance. The recognition task consisted of recognizing names (in English) contained in a database, in a speaker-independent mode. The baseline was established by selecting 250 samples from each speaker, such that an initial recognition rate of 94% was achieved.

Speech Database:

The database used in our experiments contains 739 names, spoken by five speakers (three females and two males), or a total of 3695 individual utterances. The audio samples were recorded using the handset of a fixed-line telephone, in a relatively quiet office environment. In fact, the initial (estimated) SNRs for the audio files in the database were more than 35 dB. In view of this relatively 'clean' name database, the embedded noise P'_n in (1) will discarded for the rest of our discussions. The various background noises were then digitally added to the audio files in the database to result in a given SNR.

Noise Database:

Six typical noises were used in our experiments. Five were selected from the NOISE-92 database: white noise, pink noise, brown noise, babble (cafeteria) noise, car noise (Volvo car); the remaining noise sample was recorded computer fan noise, in an office with three running computers to simulate conditions of a typical office. The Knowles EA-1842 microphone was used for recording the data, at sampling rate of 8kHz and 16 bits solution, using the amplifier of a TASCAM DA-P1 DAT recorder as the microphone pre-amplifier. To test the telephony speech recognition, all the noise types was processed by a bandpass filter (300-3400 Hz) before they were added to the audio files in the speech database, as seen in Fig. 2.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the experimental results will be presented and their implications for (a) additive noise, and (b) convolutive effects discussed.

4.1. Additive Noises

The relationship between telephony speech recognition rates and the PESQ MOS will be examined using six typical additive noise types at the SNRs of 5, 10 15, 20 and 25dB. As shown in Fig. 3(a), noises from Volvo car noise and the computer fan, at the same given SNR, have the least effects on recognition rates comparing to the other noise types. More importantly, it can be clearly seen that at the same SNR the recognition rates vary significantly with types of noises applied. As a result, SNR alone cannot be used

to predict or estimate the speech recognition performance. The corresponding PESQ MOS for the different noise types is shown in Fig. 3(b), where it can be seen that the PESQ MOS also varies with different noise types at the given SNRs. However, the almost linear trend between SNR and PESQ MOS is unmistakable. Similarly, the Volvo car and computer fan noises have relatively high PESQ MOS, i.e. better perceptual quality, which is consistent with the previous observation.

Based on the results in Figs. 3(a) and 3(b), a relationship between the recognition rates and PESQ MOS under various additive noises and at different SNRs can be established, as shown in Fig. 4. Here, the line-of-best-fit is a 4th-order polynomial fit using the criterion of minimum mean squared error. A close relationship between speech recognition rates and PESQ MOS can be observed. Furthermore, this relationship is independent of the noise type, as well as the SNR. This seems to suggest that the PESQ MOS can provide a good prediction forecast of the speech recognition rate, through the 4-order polynomial. The results in Fig. 4 also suggest that there are two regimes of operation, with relatively high recognition rates when PESQ MOS is greater than 3 and a sharp deterioration when PESQ MOS is less than 3. This can provide useful guidelines in the deployment of telephony-based recognition systems.

4.2. Convolutive Effects

In this section, we will examine the degradation in speech quality as the distance between microphone and the speech source is varied.

From Fig. 5(a), which shows the recognition rate against the source-microphone distance, it is perhaps not surprising that the recognition rates decreased as distance between the speech source and the microphone was increased: from an approximately constant level of 90% at 40cm separation, to a final recognition rate of 76% at the recording distance 80cm. The reverberations in a realistic environment is as yet not satisfactorily solved problem and can cause serious problems for speech recognizers. A similar trend can be seen when babble noise was introduced, although with the degradation appearing at a much faster rate, achieving above 85% (compared with about 90%) only in the close talking situation (10cm and 20cm), and with recognition rate of less than 10% at 80cm. It can also be seen that such trends are dependent on the noise types. The corresponding PESQ MOS for the same setting is shown in Fig. 5(b), which can be more interpreted since perceptual quality is a measure that is more easily related to.

A plot of recognition rate against PESQ MOS (shown in Fig. 6) reveals a similar relationship to that of Fig. 4. It was also observed that when the polynomial line obtained in Fig. 4 was transferred directly to Fig. 6, a similar correlation can be found. The two regions of operation are again present. This indicates that PESQ MOS has a close relationship with recognition rate and based this observation, we may tentatively conclude that the recognition rates for telephony speech may be predicted using the polynomial fit according to the PESQ MOS graph as shown in Figs. 4 and 6.

5. CONCLUSIONS

The relationship between a telephony speech recognition system (Nuance 8.0) and the ITU-T P.862 PESQ MOS has been investigated under various experimental conditions: 6 types of additive noise with different SNR, as well as at various recording distances in order to account for convolutive effects. A relationship between the subsequent decrease in speech quality and the corresponding decrease in recognition rate has been established. Our initial results indicate that 1) quantitatively, there is no systematic correlation between recognition rate and a given SNR, and, more importantly, 2) the speech recognition rates can be mapped to the PESQ MOS by a simple polynomial rule. Furthermore, this relationship seems to be consistent for both additive noise and situations where convolutive effects are present. That is, the rule is independent of the noise types as well as the SNR, with the key parameter being the perceptual quality as provided by PESQ MOS.

The main implication is that the PESQ MOS may be used to predict the likely achievable speech recognition rates (relative to some baseline performance) in a real applications. This can potentially lead to much savings in the time on data recording, model training and real-life testing, provided an estimate of the speech quality measurement is available. Future study will be continued on more complicated databases, other telephony speech recognition engines to further establish and verify such relationship systematically.

6. REFERENCES

[1] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Experiments of hmm adaptation for hands-free connected digit recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 1, May 1998, pp. 473–476.

[2] Y. Gao, B. Ramabhadran, J. Chen, H. Erdogan, and M. Picheny, "Innovative approaches for large vocabulary name recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 1, May 2001, pp. 53–56.

[3] Y. Gong, "Speech recognition in noise environments: a survey," *Speech Communications*, vol. 16, no. 3, pp. 261–291, Apr. 1995.

[4] International Telecommunication Union, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, Feb. 2001.

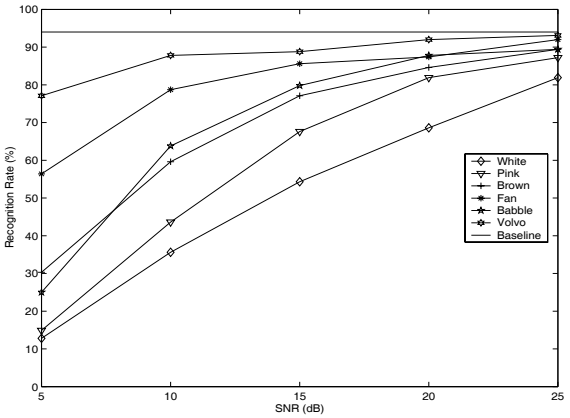
[5] W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq), the new itu standard for end-to-end speech quality assessment. Part I - time-delay compensation," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755–764, Oct. 2002.

[6] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (pesq), the new itu standard for end-to-end speech quality assessment. Part II - psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765–778, Oct. 2002.

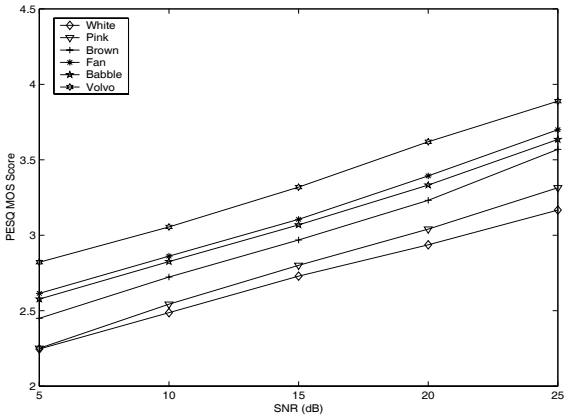
[7] International Telecommunication Union, *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*, ITU-T Recommendation P.861, Feb. 1998.

[8] ——, *Subjective performance assessment of telephone-band and wideband digital codecs*, ITU-T Recommendation P.830, Feb. 1996.

[9] J. H. L. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP-98: Int. Conf. On Spoken Language Processing*, vol. 7, Dec. 1998, pp. 2819–2822.

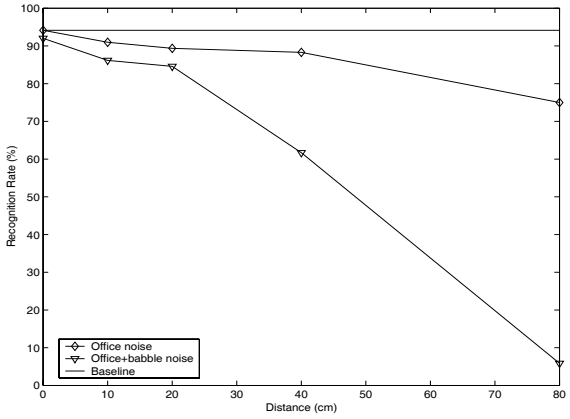


(a)

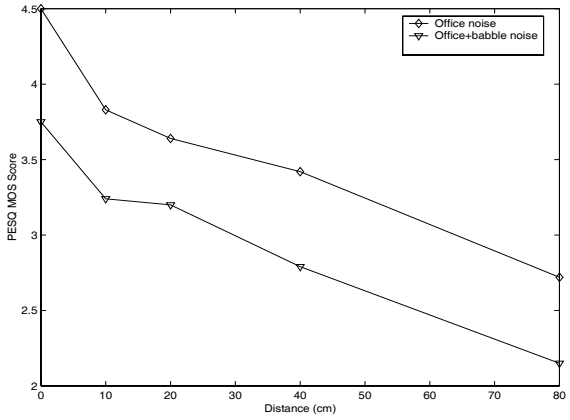


(b)

Fig. 3. The performance of speech recognition under various additive noises. (a) Speech recognition rate vs SNR; (b) PESQ MOS vs SNR.



(a)



(b)

Fig. 5. The performance of speech recognition with speech data recorded at different microphone-source distances and background noises. (a) Speech recognition rate vs. distance; (b) PESQ MOS vs distance.

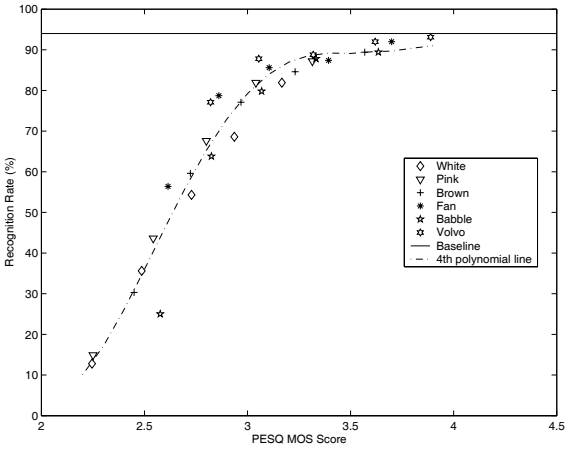


Fig. 4. Recognition rate vs PESQ MOS, under various additive noises and SNR. The 4th-order polynomial curve was empirically determined.

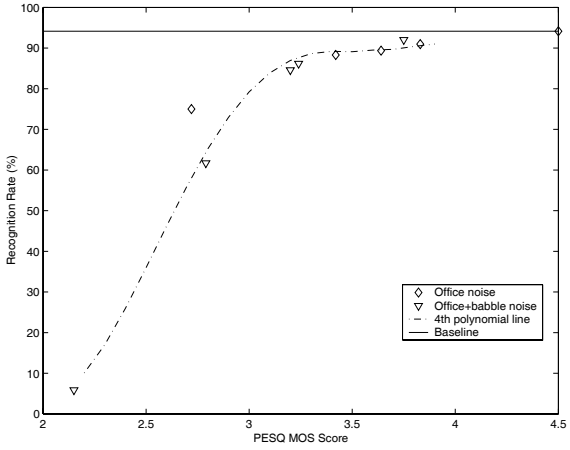


Fig. 6. Recognition rate vs PESQ MOS, at different microphone-source distances and background noises. Note the 4th-order polynomial line is the same as the one obtained in Fig. 4.