A Speech Quality and Intelligibility Assessment Project Using Google Voice

Dr. Ying Yu, University of Hartford

Dr. Ying Yu received her B.Eng. from Fudan University, Shanghai, China, in 2000. She received her M.Eng. and Ph.D. in Electrical Engineering from Brown University, R.I., USA, in 2003 and 2007, respectively. Since 2008, she has been teaching at the University of Hartford. Her current research interests are audio and speech signal processing, acoustic scene classification, speaker identification and verification, and teaching with new educational methods, including peer instruction, video games, and state-of-the-art CAD tools.

A Speech Quality and Intelligibility Assessment Project Using Google Voice

Abstract

Speech quality and intelligibility assessment is an essential topic in the area of audio and speech signal processing. It has become increasingly more important due to the rising demand in areas such as digital protocol analysis, Dolby-dts compliance analysis, audio device testing, etc.

"Audio and speech signal processing" or similar courses are widely offered as senior elective or graduate level courses in the electrical engineering curriculum. One of the biggest challenges in conducting a course project focused on speech quality and intelligibility assessment is the lack of available data from real telephony systems. In this paper, a speech quality and intelligibility assessment project using Google Voice is introduced. The voicemail feature of Google Voice allows users to download their own voicemails which are real Voice-over-Internet-Protocol (VoIP) speech. Google Voice also automatically transcribes the voicemails, providing great material for students to examine the effects of vocabulary and contextual information on the intelligibility of speech.

The Students' feedback regarding their Google Voice user experience and learning experience were collected through an exit-survey. The survey results suggested that in general the students found it convenient and easy to record their speech using Google Voice; they also agreed that using real-life data offered realistic tests of the theory, and that the automatic transcription system allowed them to investigate the performance of a real-life speech recognition system. In closing, the conclusions and future plans are presented.

Introduction

In recent years, speech and audio processing has received significant attention [1][2][3] in the engineering education society, while little has been proposed regarding topics related to speech quality and intelligibility assessment. Speech quality and intelligibility assessment has become a topic of increasing importance in research literature [4][5][6] due to the rapid development of telephony systems of various kinds of technology, VoIP being one of the fastest growing. "Audio and speech signal processing", or similar courses such as "digital speech processing", "automatic speech processing", are widely offered in modern electrical engineering curriculum. However, most of these courses focus on traditional topics, such as speech coding, speech synthesis, and speech recognition for their course projects [7][8][9]. Few has made speech quality and intelligibility assessment as the focus of a major project, even though this topic is becoming increasingly important both in research literature and in industrial practice.

At the University of Hartford, "audio and speech signal processing" was offered for the first time as an elective course for both electrical engineering senior students and graduate students during the Fall 2013 semester. There were totally 18 students in the class. The prerequisite for this course is "digital signal processing", which means all students should have basic knowledge of the sampling theory, digital filter analysis and design, z-transform, and Matlab software, etc. During the course, a project focused on speech quality and intelligibility assessment was conducted by the students with real-life VoIP data recorded by the students themselves using Google Voice. The project was designed to last three weeks: during the first week, the students recorded synchronized data using Google Voice and Matlab; during the second week, the students conducted subjective speech and intelligibility assessment; during the third week, the students conducted objective assessment and compared the results with those of the subjective assessment. The amount of work required of the students outside of class was about 2~4 hours per week.

Traditionally, one of the biggest challenges in studying speech quality and intelligibility assessment is the lack of availability of free-to-access real-life data. One would have to use simulated speech dataset by artificially corrupting the source dataset to simulate the degradation present in real-life telephony systems. Simple methods like adding acoustic background noise is easy for students to complete but far from being realistic. More sophisticated methods that offer suitable simulated effects require running multiple speech codec packages, which is far beyond the knowledge and skill levels of most students. To the author's knowledge, Google Voice is the first and only free-to-use software product available right now that allows users to record and download their voicemails without requiring any special equipment aside from a phone and a personal computer. In designing the project to require students to create their own VoIP dataset using Google Voice, the author considered the following advantages and benefits:

- Having students conduct their own data recording and construct the dataset makes them better understand the procedures of the traditional subjective speech quality and intelligibility assessment methods.
- Using real-life VoIP data by Google Voice gives students the opportunities to study some of the realistic challenges in speech quality assessment, including package losses, codec distortions, noise suppression artifacts, etc.
- The voice transcription system of Google Voice enriches the students' learning experience by relating and expanding to other aspects of speech signal processing.

The rest of the paper is organized in the following sessions: 1) project overview; 2) speech quality and intelligibility assessment; 3) creating VoIP recordings using Google Voice; 4) Google Voice automatic transcription system; 5) summary of student feedback; 6) conclusions and future plans.

Project Overview

The speech quality and intelligibility assessment project introduced in this paper includes three phases. In phase one, the students created a speech data bank using their own voice recordings. Each student created two sets of recordings. For the first set of recordings, they read the famous "rainbow passage" [10] using continuous speech. For the second set of recordings, they read a randomly selected group of words in the classic Modified Rhyme Test (MRT) [11] using discrete speech. For each reading, they simultaneously recorded their speech using a close-talking microphone on their own PC, and at the same time recorded their speech with Google Voice's [12] voicemail system through a phone. After uploading their own recordings onto a commonly shared webpage (e.g. Blackboard's 'file exchange' page), all students had access to each other's recordings. After all the speech recordings were collected, the complete speech database comprised of 2 sets of synchronized recordings from each of the 18 students.

In phase two, the students subjectively assessed the speech quality and speech intelligibility of each recording using the Mean Opinion Score (MOS) [13] [14] system and modified rhyme listening tests.

In phase three, the students assessed the speech quality of the recordings objectively using the Perceptual Evaluation of Speech Quality (PESQ) [15] [16] algorithm which is a legacy industrial standard for objective voice quality testing.

Traditionally, one of the biggest challenges in studying speech quality and intelligibility assessment is the lack of availability of free-to-access real-life data. The voicemail feature of Google Voice allows students to download their own voicemails which are real VoIP speech. The voicemails are sampled at 11025Hz and compressed as MPEG audio layer III (mp3). Google

Voice also automatically transcribes the voicemails. There are many ways to use this feature to enhance students' learning experience. For example, when the students recorded different types of speech, continuous vs. discrete speech, they observed dramatically different speech recognition performances. This offered them a great opportunity to investigate the performance of a state-of-the-art speech recognition system.

Speech Quality and Intelligibility Assessment

A classic subjective measurement of speech quality is Mean Opinion Score (MOS). This usually involves a group of listeners, who will listen to the audio recordings under evaluation and then rate the recordings according to the scale shown in Table 1. The MOS score of a certain recording will be the average of the scores reported by all the listeners. The MOS test is standardized by the International Telecommunications Union (ITU) in recommendation P.800. A great reading material for students to understand the procedures and requirements of the most important subjective and objective speech quality measurements is the paper "speech quality assessment" by Loizou [17].

Speech intelligibility is best measured subjectively by a panel of listeners, who will try to identify words, phrases or sentences. One of the traditional tests is the modified rhyme test (MRT). It is designed by linguistic specialists in 1960's and still used by audio and speech researchers today [18]. MRT asks listeners to listen and select one of six words, half of which differing by the initial consonants, and another half differing by the final initial consonants. The overall averaged recognition rate, the total number of correctly identified words divided by the total number of words tested, will be the measurement for speech intelligibility. An example of a six-word MRT group is listed in Table 2. A total of fifty such groups comprise the test used in the project.

Table 1: Description of MOS Scores and Corresponding Impairment Levels

Score	Description	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 2: an example of a six-word MRT group

game	came
fame	name
tame	same

Subjective measurements are very time consuming, lack repeatability, and as the number of listeners increases, they also become very expensive. This makes objective methods to measure speech quality very desirable in today's fast growing telephony market. PESQ is the most commonly used objective measurement system. It is also a global industry standard that any newly developed method compares to. PESQ involves a comparison between a degraded speech signal and a reference speech signal to predict the MOS value. That is also the reason why the students need to conduct recordings in synchronized pairs. In the project's phase three, when they were conducting objective assessment using the PESQ algorithm, they used the Matlab recorded "clean" speech recordings as the reference signals and the Google Voice recorded VoIP speech recordings as the degraded signals whose PESQ values were to be calculated .

Creating VoIP Recordings using Google Voice

To create VoIP recordings along with synchronized reference recordings at the same time, the students followed the instructions below:

1. Create a Google Voice account (must be a "Google number" account) and call the number with a phone and leave a voicemail. The voicemail message can be downloaded in mp3 format.

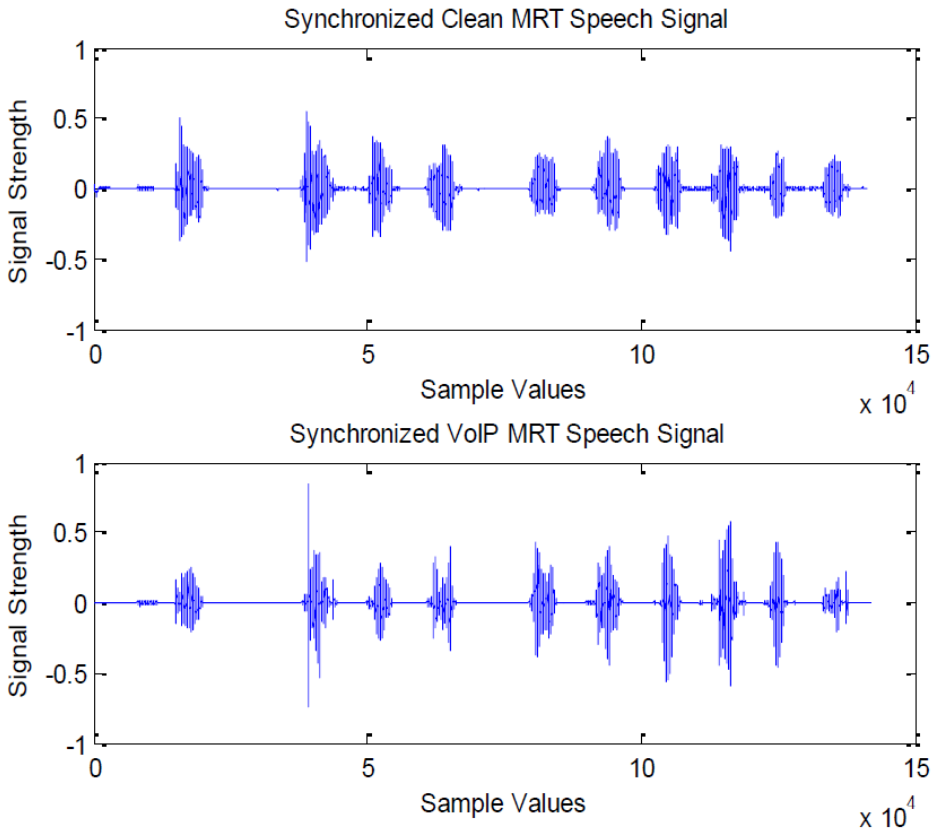

Figure 1: Synchronized "clean" and VoIP speech recordings of MRT by a student

2. To record a Google voicemail and a clean Matlab recording at the same time requires multi-tasking on the students' part. Use a phone to call the Google voice number first, and wait for the voicemail to pick up. Activate the "wavrecord" recording command in Matlab and then start reading the rainbow passage or MRT test. Make sure your mouth is positioned properly for both the close-talking microphone connected to the computer and the microphone of the phone. When finished reading, hang up the phone and wait for wavrecord to be finished as well.

3. Download the voicemails as mp3 files and then read them into Matlab using function mp3read.m (students can download mp3read.m from Mathwork's "Matlab central" file exchange website) and convert them to wave files.

4. Synchronized the VoIP and Matlab recordings by using either the cross-correlation method or visual inspection of the time-domain wave plots.

An example of a pair of synchronized "clean" speech recording and VoIP speech recording is shown in Figure 1 on the previous page. The recordings are of a student reading words from MRT using discrete speech. The VoIP speech can be degraded in many ways due to packet loss, delay, jitter, codec, etc. [19] Another example in Figure 2 shows the significant difference in the silence period of a reference signal vs. that of the corresponding VoIP degraded signal.

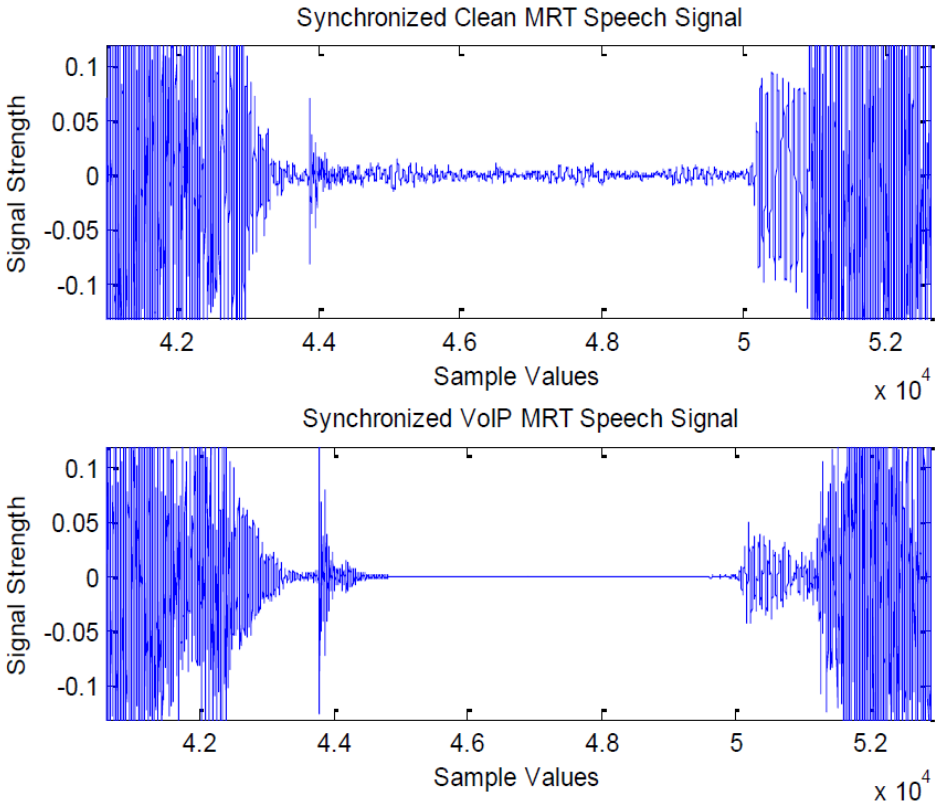

Figure 2: Detailed view of silence periods of a pair of "clearn" vs. VoIP recordings by a student

In Figure 2, the complete absence of signal in the VoIP recording during the silence period is very striking, unlike other telephony systems. This is due to a voice detection algorithm commonly used in VoIP systems. Since PESQ was not developed for VoIP systems, its values could be misleading for such situations where silence dominates. The phenomenon was observed by the students and offered great learning opportunities where the students were able to learn about the limitations of the PESQ algorithm. According to the students' feedback from the exit-survey results, the students felt that using real-life data like the voicemail recordings of Google Voice in the project offered realistic tests of the theory.

Google Voice Automatic Transcription System

In the project, the students read and recorded with Google Voice's voicemail: the first paragraph of the "rainbow" passage in continuous speech; 50 words from the MRT in discrete speech. They then used the voicemail transcription results to calculate the correct rate for each recording.

Table 3 is the result summary by a student for the "rainbow passage" read in continuous speech. Table 4 is the result summary by the same student for the "MRT" read in discrete speech.

Table 3: Result summary of Google Voice transcription system with continuous reading of the "rainbow" passage by a student

Trial #	Time (s)	Correct words	Success Rate
1	38	84	85.71%
2	38	89	90.82%
3	36	81	82.65%
Average	37.33	84.67	86.39%

Table 4: Result summary of Google Voice transcription system with discrete reading of the MRT by a student

Trial #	Time (s)	Correct words	Success Rate
1	49	21	42.00%
2	53	21	42.00%
3	47	16	32.00%
Average	49.67	19.33	38.67%

The dramatically different "success rates" in Table 3 and Table 4 suggest that the speech recognition system of Google Voice is designed for continuous speech and uses contextual information. That is why MRT, consisting of all unconnected random words, received much worse results. According to the students' feedback through the exit-surveys, the students felt that

the Google Voicemail automatic transcription system allowed them to investigate the performance of a real-life speech recognition system.

Student Feedback Summary

An anonymous exit-survey was conducted at the end of the semester. The exit-survey contained six questions. A total of 15 students responded. The first question asked the students what type of computers they used. The answers showed that 12 students used their own laptop with Microsoft Windows OS; 2 students used the university desktops, which also have Microsoft Windows OS; 1 student used his/her own Mac laptop. Questions 2~5 were Likert opinion questions with the following scale: the value 5 indicating strong agreement, 4 moderate agreement, 3 neutral, 2 moderate disagreement, and 1 indicating strong disagreement. The question descriptions, Likert score average values, and standard deviations are summarized in Table 5.

Table 5: Summary of Likert Opinion Questions in the Exit-Survey

	Likert Opinion Question Description	Likert Score Average	Standard Deviation
Q2	Using Google Voice to record my speech is convenient and easy.	4.13	0.64
Q3	The instructions in the handout are helpful in terms of getting me started to use Google Voice.	4.53	0.64
Q4	Using real-life data like the voicemail recordings of Google Voice in the project offers realistic tests of the theory in terms of speech quality and speech intelligibility.	4.53	0.52
Q5	The Google Voicemail automatic transcription system allows me to investigate the performance of a real-life speech recognition system.	4.13	0.99

The last exit-survey question was an open-ended one where the students were asked to leave any comments or suggestions regarding how to improve the project "speech quality and speech intelligibility assessment using Google Voice". A total of 11 students responded. Most expressed their appreciations: ‘The project was very good and it helped us to test the real life speech recognition system … it was very informative and we had great experience playing with google voice and testing its ability to recognize speech”; “Overall, Google voice is a good tool to detect speech quality and speech intelligibility assessment”; “The Google Voice aspect of the project is sufficient and relatively easy to complete”; “Google voice is a neat program offered by google and it shows where speech dsp is at these days”.

Several gave suggestions: “I'd like to know more about how google voice works”; “adjust the recording devices to the best status and try to use the same device for everyone's recordings,

keep everything at the same quality"; "it might be worthwhile to use a separate voice recognition software to compare the results with Google"; "maybe students need a sample recording and they can follow the sample recording to do their own recordings. That might keep everyone has the same standard"**.** One student expressed concern that his/her Mac laptop was not readily recognizing ".wav" files. One student expressed concern that s/he didn't want his/her number to be used on a marketing list.

Conclusions and Future Work

In summary, the author believes that the speech quality and intelligibility assessment project using Google Voice was a success. The Students' feedback suggested that in general the students found it convenient and easy to record their speech using Google Voice; they also agreed that using real-life data offered realistic tests of the theory, and that the automatic transcription system allowed them to investigate the performance of a real-life speech recognition system. In the future offerings of the audio and speech signal processing course, the author plans to improve the project by: 1) offering students a trouble-shooting guide regarding how to use Google Voice; 2) providing a reading sample recording so that students have a better understanding of the pace for reading MRT; 3) providing more uniform recording devices so that the qualities of recordings are more consistent.

References:

[1] T. Ogunfunmi, "Pedagogy of a course in speech coding and voice-over-IP", ASEE 2008 Annual Conference Proceedings, AC2008-2673
[2] B. Barkana, "A graduate level course: audio processing laboratory", ASEE 2010 Annual Conference Proceedings, AC2010-1594
[3] V. Kepuska, M. Patal, N. Rogers, "A Matlab tool for speech processing, analysis and recognition: SAR-Lab", ASEE 2006 Annual Conference Proceedings, AC2006-472
[4] T. Falk, W. Chan, "Performance study of objective speech quality measurement for modern wireless-VoIP communications", EURASIP Journal on Audio, Speech, and Music Processing, Volume: Jan. 2009, Article No. 12, doi: 10.1155/2009/104382
[5] S. Moller, W. Chan, N. Cote, T. Falk, "Speech quality estimation: models and trends", IEEE Signal Processing Magazine, Volume: 28, Issue: 6, Nov. 2011, Pages: 18-28, doi: 10.1109/MSP.2011.942469
[6] J. Ma, P. Loizou, "SNR loss: a new objective measure for predicting the intelligibility of noise-suppressed speech", Speech Communication, Volume 53, Issue 3, March 2011, Pages 340-354
[7] J. G. Harris, course webpage for "Automatic Speech Processing", http://www.cnel.ufl.edu/hybrid/courses/EEL6586/
[8] L. R. Rabiner, course webpage for "Digital Speech Processing", http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20course.html
[9] M. Brookes, course webpage for "Speech Processing", http://www.ee.ic.ac.uk/hp/staff/dmb/courses/speech/speech.htm
[10] G. Fairbanks, "Voice and Articulation Drillbook", Joanna Colter Books; 2nd edition, 1960
[11] A. House, C. Williams, M. Hecker, K. Kryter, "Articulation testing methods: Consonantal differentiation with a closed response set", Journal of the Acoustical Society of America, 1965, JASA 37: 158-166
[12] www.google.com/voice
[13] "Methods for Subjective Determination of Transmission Quality, ITU, Rec. P.800", Int. Telecommun. Union, Aug. 1996.
[14] http://www.itu.int/rec/T-REC-P.800.1-200303-S/en

[15] “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 Rec. P.862”, International Telecommunication Union, Feb. 2001.
[16] http://www.itu.int/rec/T-REC-P.862/en
[17] P. C. Loizou, “Speech Quality Assessment”, Multimedia Analysis, Processing and Communications: Studies in Computational Intelligence Volume 346, 2011, pp 623-654
[18] http://www.pscr.gov/projects/audio_quality/mrt_library/mrt_library1.php
[19] J. Davidson, J. F. Peters, M. Bhatia, S. Kalidindi, S. Mukherjee, “Voice-over-IP Fundamentals (Second Edition)”, Cisco Press, 2006