ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(24), DOI: 10.17485/ijst/2015/v8i24/83106, September 2015* ISSN (Online) : 0974-5645

Intentional Estimation of Analytical Oration

D. Gayathri, S. Sathya Devi, M. Niranjana Priyadarshini and M. Divya

Electronics and Communication Engineering PSNA College of Engineering and Technology
Dindigul – 624622, Tamil Nadu, India; deivagayathri@gmail.com[3]priyadarshinim89@gmail.com

Abstract

The speech quality assessment is categorized into two: subjective evaluation and objective evaluation. The subjective evaluation is mainly based upon the perceptual quality of the artificial speech which is carried out by means of listener's rating. But, here the drastic variation at the points of concatenation is not taken into account. Hence, objective measures are the right alternative for evaluation. Here, the time domain parameters like energy, intonation and duration are analyzed for the synthetic speech. The rating is given on the scale of 1 for the variations and this is deduced from the subjective evaluation results, as the overall quality cannot be purely based on objective measures. The synthetic speech is synthesized by a phoneme based Unit Selection Synthesizer (USS) consisting of three hours speech corpus. The objective evaluation is assigned a value of 1, among which more priority is given to energy as 0.7 in 1 and less priority is given to duration as 0.3 in 1, based on its contribution to the quality and variations. Intonation is not considered, as its characteristics are not revealed at phoneme level. Finally the objective MOS 0.704 is reduced from the actual MOS 2.75 and the overall rating obtained is 2.046.

Keywords: Duration, Intonation, Objective Measures, Short Time Energy, Subjective Measures, Unit Selection Synthesizer

1. Introduction

Speech is the phonetic combination of vowel and consonant speech sound units. A speech synthesizer is the system that converts the given input text into an artificial speech. Speech synthesis system is broadly classified into concatenative speech synthesizer and statistical parametric speech synthesizer. Among the concatenative synthesizer, Unit Selection Synthesizer (USS) is commonly used in synthesizing speech.

The speech quality assessment is categorized into two: subjective evaluation and objective evaluation. The need for the speech evaluation is for analyzing the intelligibility of artificial speech compared to that of original speech. The subjective evaluation is performed by providing a comparative rating for the artificial speech. But, here the drastic variation at the points of concatenation is not taken into account. So, subjective evaluation is overcome by objective evaluation.

Here, the synthesized speech waveforms are provided for analysis, where the time domain parameters like energy, intonation and duration are analyzed. The analysis is performed by comparing the original speech and the artificial speech, synthesized for the same sentence. After analyzing, the rating is given on the scale of 1 for the variations and this is added to the subjective evaluation results, as the overall quality cannot be purely based on objective measures. Here, the synthetic speech produced by a phoneme based USS using three hours speech corpus is analyzed objectively.

2. Subjective Measures

The subjective evaluation is mainly based upon the perceptual quality of the artificial speech. The audio or speech files that are to be evaluated are played to all the speakers simultaneously by using a loud speaker in a noise free environment. The synthetic speech is

*Author for correspondence

evaluated by listeners as per their own opinion. So the quality of artificial speech cannot be well defined as the drastic variation at the points of concatenation is not taken into account. The following methods are used in the subjective evaluation of artificial or synthetic speech: MOS, CMOS, DMOS, DRT and Audio Visual Assessment.

1.1 MOS

Mean Opinion Score (MOS) is probably the most widely used[16] and simplest method to evaluate speech quality. It is also suitable for overall evaluation of synthetic speech. In voice and video communication, quality usually dictates whether the experience is a good or bad one.

2.2 CMOS

The evaluators listen to original and artificial speech, and make their assessment by comparing the two. However, the ordering of the two speeches is changed randomly. The evaluator's opinions of the second speech sample compared with the first speech sample. The average score from a sufficient number of evaluators is called the CMOS score.

2.3 DMOS

The evaluators compare the speech to be assessed with original speech to assess the degree of degradation in the speech samples. Specifically, they first listen to the original speech and then listen to the speech to be assessed after a delay of 0.5–1 second, in order to avoid the influence of first played speech over the second speech.

2.4 DRT (Diagnostic Rhyme Test)

In DRT, respondents hear a word and choose its equivalent from two visually presented words. The two words differ only in their initial and the consonants differ only in a single distinctive acoustic phonetic feature[7].

2.5 Audio Visual Assessment

The intelligibility of audio visual speech can be evaluated as normal speech[12]. It is feasible to compare the results of other combination of natural and synthetic speech. Intelligibility increases with facial information. This method of subjective evaluation increases the capacities of computer graphics.

In subjective evaluation, the variations at the point where the phonemes join cannot be seen visually i.e. the concatenation points are not seen. The quality is partially rated as it depends on individual opinion[6].

3. Unit Selection Synthesizer

Unit selection synthesizer is the most commonly used concatenative synthesizer[10]. Unit selection synthesis uses large database of recorded speech as shown in Figure 1. During database development, each recorded utterance is segmented into individual phonemes, half-phones, syllable or other sound units like CV. The USS can be developed for various languages during which a transliteration is required[3], using different size of speech corpus. With increase in size of corpus, the quality of speech will also increase. Unit selection synthesizer generates speech by selecting proper units from a speech corpus. It contains three main parts such as, text analysis, selection of units, speech generation module.

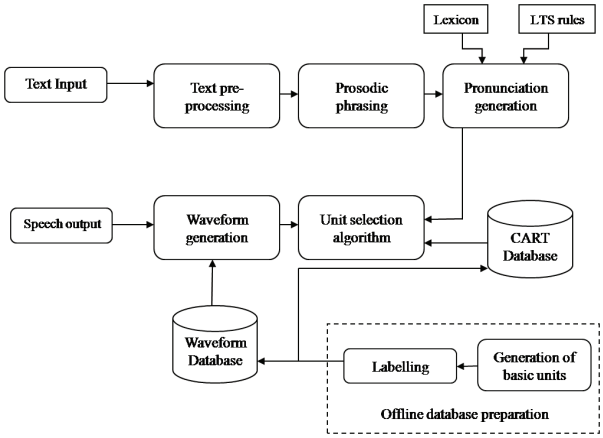


Figure 1. Unit selection synthesizer.

3.1 Speech Corpus

The speech corpus consists of speech files and its transcriptions[20] that are recorded in an anechoic chamber. It can also be recorded in an echo free laboratory environment using microphone and audio mixers. Mixers are used for reducing noise. The sampling rate while recording is set as 16KHz. These recorded files are stored as wave files. The transcriptions for these files are obtained by means of segmentation. The transcriptions are called label files. It consists of each phoneme and its start and end time.

3.2 Text Analysis

It is used to analyze and normalize the text based on LTS (letter to sound) rules[21]. During normalization the

punctuation marks are removed for the given input text and the input text are separated in to the required text segments. Ex: the word ammA is separated in to /a/, /m/, /m/, /A/ for phoneme based system, as /am/, /mA/ for CV based system. Here, the most appropriate units are selected for each text segment by means of a unit selection algorithm. In this case, appropriate phoneme units are selected as it is phoneme based USS. These speech waveforms for each phoneme unit are joined or concatenated together and the entire waveform for the given input text is obtained.

4. Time Domain Parameters

The time domain methods is the set of processing techniques that involves the waveform of speech signal directly. The representations of speech signal in terms of time domain measurements[8] include: Short Time Energy (STE), Duration and Intonation.

4.1 Short Time Energy

Speech is described as a slowly time-varying or locally stationary process that contains many frequencies. Hence the speech waveform is known to be quasi-stationary. Because of the slowly varying nature of the speech signal, it is to be processed as blocks called frames. This leads to the basic principle of short-time energy analysis. The unit for energy is joules.

4.1.1 Energy of Sound Units

The regions in speech are broadly divided into two. They are: voiced and unvoiced. The energy levels of each sound unit vary. The voiced part[1,4] of the speech has high energy because of its periodicity and the unvoiced part of speech has low energy. The amplitude variations of the speech signal depend upon the short-time energy. The short time energy of the speech is defined as[15]:

$$En = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]\wedge 2 \quad (1)$$

Where, E is the energy
n is the frame number
m is the no. of samples
x(m) is the speech signal
w(n-m) is the window function

4.1.2 Hamming Window

Hamming window is chosen as the rectangular window has sharp edge and hamming window has larger height of side lobes. The bandwidth of hamming window is about twice the bandwidth of rectangular window of same length. Thus, the hamming window function is selected for minimizing the height of nearest side lobes and has smooth rolling edges as given in Figure 2. Hamming window is also one period of raised cosine and it is much better than hamming and rectangular window functions. The optimized side lobes nearest to the main lobe occupy a smaller frequency interval about main lobe. The samples in a frame of speech varies from m=0 to m=N-1.

$$H_H[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right), 0 \le n \le L-1 \\ 0, otherwise \end{cases} \quad (2)$$

Where, H_H (n) is the window function
L is the no. of samples

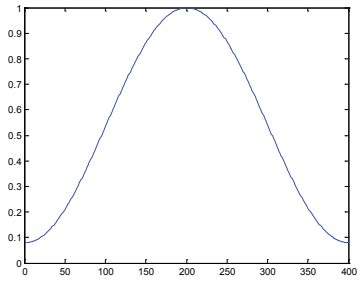


Figure 2. Hamming window.

1.1.3 Window Frame Size

The window frame size used in determining the short time energy of the input speech is 400 samples. The mathematical reason is expressed as follows for defining the frame length:

$$\begin{aligned} F &= Fs/N \\ 1/T &= Fs/N \\ N &= T^*Fs = 25ms^*16kHz \\ N &= 400 \end{aligned} \quad (3)$$

The reason for choosing the pitch period to be 25 ms [14] is that, it is necessary to consider the pitch periods of both male and female. Irrespective of the gender, the period chosen should accommodate at least a few formant frequencies facilitating its processing. Moreover, the sampling rate while building the speech corpus is set to be 16 KHz, in order to satisfy the Nyquist criteria, as the maximum frequency occupied by fricatives are up to 8 KHz.

4.1.4 Energy Weightage

As short-time energy plays a vital role in the speech quality evaluation and as more variations were found in it, a larger weight age of rating value is provided in OMOS rating. The overall OMOS value is taken as 1 and due to higher priority to energy, it is assigned an OMOS value of 0.7 in 1for the all phoneme classifications.

4.2 Duration

Duration is defined as the length of time interval of each individual sound unit. The quality of speech is affected due to the time interval of the phoneme. Here, the USS selects the required units of different time span. Generally, the time taken for the voiced sound units is longer in duration than the unvoiced sound units. Thus the duration of each phone has its own uniqueness of duration as in energy.

4.2.1 Segmentation

Segmentation is the process required for labelling a wave file. The segmentation is used for identifying the individual characteristics of each phoneme. The segmentation is done for both original and artificial speech signals in which the sentences are phonetically balanced. The segmentation process can be done by following two methods: Manual Segmentation and Forced Viterbi alignment.

4.2.2 Manual Segmentation

Manual segmentation is carried out by using HTK (Hidden Markov Model Tool Kit) transcription as segmentation configuration as in Figure 3. The selected sentence are segmented and labeled as phoneme units. The main drawback in manual segmentation is that it is difficult to segment the phonemes in synthetic speech as it has influence of neighbouring phoneme or some other factors like conjunction that affects the quality of synthetic speech. It is also time consuming and difficult. Hence, an automated segmentation is required. As HTK transcription is used for calculating duration, the durations are obtained in terms of 100 of nano seconds.

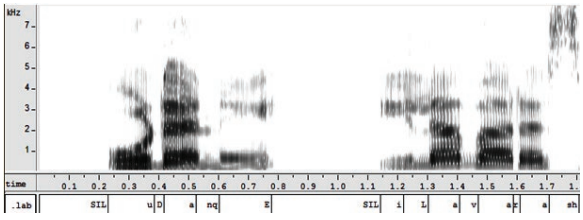


Figure 3. Segmentation using HTK transcription.

4.2.3 Forced Viterbi Alignment

Forced Viterbi alignment used to segment the wave files and provide labels which are required for analyzing time-domain parameters. Here, the voice data collected for three hours corpus are obtained. The forced Viterbi alignment is done by using HTK transcription. This requires dictionary, Master Label File (MLF), configuration files and HMM re-estimated models. The label files help in matching the text with the speech waveform and provide information about the occurrence of the speech units. Hence, the recorded speech needs to be segmented. The Master Label File contains the phoneme level transcription of the input data to be segmented. In the MLF, silences between words are introduced for better results. In order to create the MLF, orthography to phoneme conversion is carried out for the text data and to sort the phonemes phone list is used. Finally MLF is corrected manually for voiced and unvoiced pronunciations of various sounds (th, dh, T, D, k, g, h, p, b etc). Initially, by means of manual segmentation, models are trained for all the phoneme units and using these models forced Viterbi alignment is carried out where, the exact transcription of the speech file to be segmented is directly given to the search engine of a speech recognition system. The final result is a set of label files for the given speech files[22].

4.2.4 Duration Weightage

As the duration of phoneme plays a vital role in the speech quality evaluation, a weightage of 0.3 in 1 is provided as OMOS rating. The weightage is not as high as the weightage that was given to STE, as the duration does not affect the perceptual quality as that of a STE and less variation were found.

5. Results and Analysis

The analysis of energy and duration for the phonetically balanced sentences (sentences containing all phonemes) was mainly concentrated in places where there were perceptual perturbances. These perturbances were due to the inappropriate unit selection. In terms of energy, units with low energy were chosen in places where units with high energy is required (as in the middle of a word) or vice versa. In a word, the starting and ending energy are comparatively lower than the energy in the middle of the word. In terms of duration, units with longer duration are chosen in places where units with shorter duration are required and vice versa.

5.1 Steps to analyse Short Time Energy

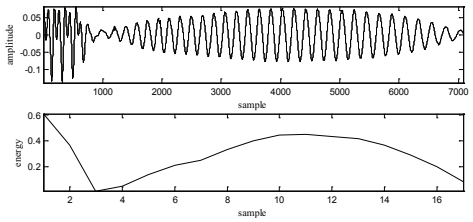
To analyze the energy of the phoneme, the following steps are performed:

- The original speech is recorded in an echo free laboratory environment or chosen from the database.
- The artificial speech is synthesized using the chosen speech synthesizer.
- For analyzing the concatenative points, the part of the speech file, containing the phoneme before and after the concatenation points are saved as separate wave files and analyzed.
- The window function and frame size are determined.
- These windows frames are used for determining the short-time energy completely.
- The energy is analyzed and compared for both original and artificial speech the energy of artificial speech is determined for the words in which the sonic glitches and some effects occurs.

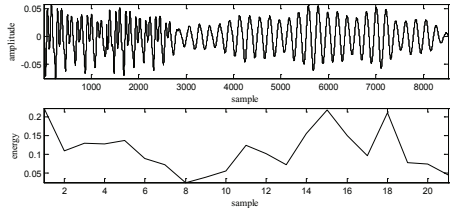
5.2 Analysis of STE

The following results have been analyzed from the variations in the energy between the original speech and artificial speech.

There is some possibility of missing short duration consonants as seen in Figure 4.



(a)



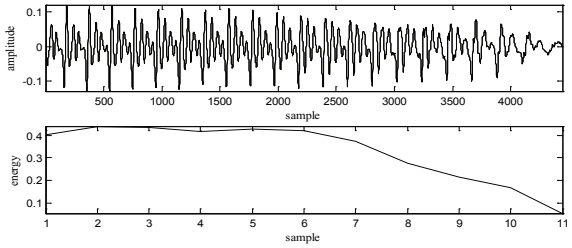
(b)

Figure 4. **(a)** Natural speech. **(b)** Synthetic speech.

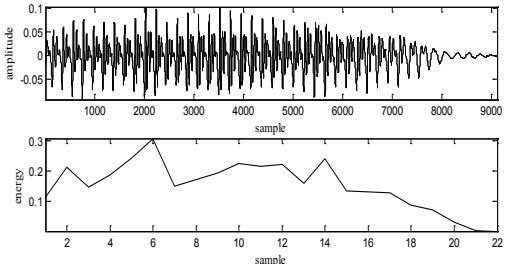
The other analysis results are that,

- There occurs some noise or influence of some other phoneme in the place of silence
- The duration of phoneme varies between original and artificial speech
- Synthesizer selects other phoneme instead of the required phoneme
- The energy of artificial speech is narrowed while the energy of the original speech is distributed
- The energy completely alters between original and artificial speech for the similar phoneme.

As seen in Figure 5. There is completely different rise and fall of energy for same phoneme



(a)



(b)

Figure 5. **(a)** Natural speech and **(b)** Synthetic speech.

5.3 Steps to Analyse Duration

Table 1. Duration variations for various phonemes

Phoneme	Original	Artificial
A	838329	600000
M	1007608	1200000
A	1500000	1100000
R	306313	300000
N	999546	800000
A	2000000	1500000
Ch	300000	1200000
Nq	3400000	1500000
Ai	1700000	2900000
E	1700000	1500000

To analyze the duration the phoneme, the following steps are performed:

- The original speech is recorded in an echo free laboratory environment or chosen from the database.
- The artificial speech is synthesized using chosen speech synthesizer.
- Segmentation is carried out for the phonetically balanced sentences that are to be analyzed.
- The phonetically segmented wave files are evaluated in duration script using shell program.
- Then, the duration is determined for both original and artificial speech. The values are compared and ratings are provided.

5.3.1 Analysis of Duration

The duration of short phoneme occupies more time duration due to the influence of neighbouring phoneme as given in Table 1. The duration of long phoneme gets very less duration due to the improper selection of units. The time domain methods is the set of processing techniques that involves the waveform of speech signal directly. The representations of speech signal in terms of time domain measurements[8] include: Short Time Energy (STE), Duration and Intonation

5.3.2 Intonation

Intonation is neglected due to the reason that, it cannot be revealed in smaller units[2] like phoneme. It can be analyzed for longer units like syllable.

5.3.3 Overall Rating

The overall rating is provided for the analyzed time domain parameters for Short Time Energy and Duration. More weightage is given to energy (0.7 out of 1) and less weightage is given for duration (0.3 out of 1) whereas intonation is neglected. The OMOS is given in Table 2. This objective rating is combined with subjective rating finally this provides the overall rating for the quality of synthetic speech.

$$\begin{aligned}\text{Overall rating} &= \text{Subjective MOS} - \text{Objective MOS}\\ &= 2.75 - 0.704\\ &= 2.046\end{aligned}$$

Thus, the overall rating obtained is 2.046 out of 6 for the quality assessment of synthetic speech.

Table 2. Objective rating

Phoneme	Energy	Duration
Vowels	0.155	0.06
Semivowels	0.125	0.018
Voiced consonants	0.125	0.066
Unvoiced consonants	0.107	0.048

6. Conclusions

Thus, the set of data that are collected are synthesized as waveforms. The energy and duration analysis are performed for both original and synthetic speech waveforms. Based on the observation, the objective MOS rating is provided for the synthetic speech to improve the quality of speech. Finally, the subjective rating is combined with the objective rating thus giving the overall rating. An evaluation cannot be purely based on objective measures, as it only considers the variations in parameters and does not consider the perceptual quality. Similarly, an evaluation cannot be purely based on subjective measures, as it only considers the perceptual quality and does not consider the drastic changes in parameters. Therefore it is concluded that an evaluation technique, which uses both subjective and objective evaluation is more effective than using a purely subjective or objective measures based technique.

7. Future Work

With the use of other longer units such as syllables, words etc, the energy, duration and other time domain parameters like pitch, zero-crossing rate and auto-correlation functions can be analyzed which helps in improvising the quality of synthetic speech[4,5]. Even spectral parameters can be used for evaluating speech. Apart from using such evaluation techniques in synthetic speech evaluation, it can also be used for other audio quality evaluation in various domains.

8. References

1. Acoustics (online). (Accessed on 2014 Mar 7) Available from: http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap10.html
2. Ahmadi S, Spanias AS. Cepstrum-Based Pitch Detection using a New Statistical V/UV Classification Algorithm, IEEE Trans. Speech Audio Processing. 1999; 7(3):333-8.

3. Avinash Chopde, ITRANS (online). (accessed on 2014 Feb 9) Available from: http://www.aczoom.com/itrans/html/tamil/node5.htm
4. Bachu RG, Kopparthi S, Adapa B, Barkana BD. Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal, in Advanced Techniques in Computing Sciences and Software Engineering, Springer Netherlands; 2010.
5. Valentini-Botinhao C, Yamagishi J, King S. Can Objective Measures Predict the Intelligibility of Modified HMM-based Synthetic Speech in Noise?, in Proc. INTERSPEECH, Florence, Italy; 2011. p. 1837–40.
6. Hirst D, Albert,Rilliard, Auberge V, Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis, in Proc. SSW3, NSW, Australia; 2011. p. 1– 4.
7. Keller E, Keller BZ (online). (accessed on 2014 Jan 4) Available from: http://www.uniger.ch/BrigitteZellnerKeller/Brigitte_ZellnerKeller_files/Publications/Keller-ZellnerKeller-00-Phonetician81.pdf
8. Hog H. Basic Parameters in Speech Processing The Need for Evaluation, © [CiteSeer]. doi: 10.1.1.300.6548, 2010.
9. Janne Pylkkonen,Phone Duration Modeling Techniques in Continuous Speech Recognition, Dept Information and Computer Science. Helsinki University, Otaniemi, Sci. Final Rep; 2004.
10. Gros JZ, Zgance M. An Efficient Unit-selection Method for Concatenative Text-to-speech Synthesis Systems, Journal of Computing and Information Technology, CIT. 2008; 16(1): 69–78.
11. Yamagishi J (online). (accessed on 2014 Jan 15) Available from: http://mi.eng.cam.ac.uk/mi/pub/Main/SeminarsSpeech/Cam-SpeechSynthesis-Seminar.pdf
12. K.Kondo, "Subjective Quality Measurement of Speech", in Signals and Communication Technology, Springer; 2012.
13. Oiler DK. The effect of position in utterance on speech segment duration in English, © [The Journal of the Acoustical Society of America]. doi: 10.1121/1.1914393, 1973.
14. Paliwa KK, Lyons JG, Kamil K. Wojcicki, Preference for 20-40 ms window duration in speech analysis, in Proc. ICSPCS, Gold Coast, QLD. 2010; 1–4.
15. Rabiner LR, Schafer RW. Introduction to Digital Speech Processing, in Foundations and Trends in Signal Processing. now publishers Inc. 2007; 1(1–2): .
16. Mean Opinion Score (online). (accessed on Jan 6.) Available from: http://en.wikipedia.org/wiki/Mean_opinion_score
17. Dong M, Lua K-T, Li H. Unit selection-based speech synthesis approach for mandarin Chinese, International Journal of Asian Language Processing. 2005; 21:45–55,.
18. Sarathy KP, Ramakrishnan AG. Text to speech synthesis system for mobile applications, in Proc. WISP, IIT Guwahati. 2007; 74-7..
19. Hendriks RC, Heusdens R, Jensen J. Adaptive time segmentation for improved speech enhancement, IEEE Transactions on Audio, Speech and Lang. Processing. 2006; 14(6):2064–74.
20. Sangeetha J, Jothilakshmi S, Sindhuja S, Ramalingam V. Text to Speech Synthesis System for Tamil, Proc. ICISC, India; .2013.
21. King S. Introduction to Speech Synthesis, Indian Academy of Sciences. 2011; 36(5):. 837–52.
22. Andersen V Speech Coding and Recognition, IT University of Copenhagen, Final Rep; 2005.
23. Mohan V. Analysis and Synthesis of speech using MATLAB, International Journal of Advancements in Research and Technology. 2013; 2(5):373–82.
24. Hu Y, Philipos, Loizou C. Evaluation of Objective Quality Measures for Speech Enhancemen, Proceedings. Audio, speech, and language processing, 2008; 16(1): 229–37.