Performance Evaluation of Speech Synthesis Techniques for English Language

Sangramsing N. Kayte, Monica Mundada, Santosh Gaikwad and Bharti Gawali

Abstract The conversion of text to synthetic production of speech is known as text-to-speech synthesis (TTS). This can be achieved by the method of concatenative speech synthesis (CSS) and hidden Markov model techniques. Quality is the important paradigm for the artificial speech produced. The study involves the comparative analysis for quality of speech synthesis using hidden Markov model and unit selection approach. The quality of synthesized speech is evaluated with the two methods, i.e., subjective measurement using mean opinion score and objective measurement based on mean square score and peak signal-to-noise ratio (PSNR). Mel-frequency cepstral coefficient features are also extracted for synthesized speech. The experimental analysis shows that unit selection method results in better synthesized voice than hidden Markov model.

Keywords TTS · MOS · HMM · Unit selection · Mean · Variance · MSE · PSNR

1 Introduction

A speech synthesis system is a computer-based system that produce speech automatically, with the conversion steps of grapheme-to-phoneme transcription of the sentences with the inclusion of prosodic features. The synthetic speech is generated

S.N. Kayte (✉) · Monica Mundada · Santosh Gaikwad · Bharti Gawali
Department of Computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India
e-mail: bsangramsing@gmail.com

Monica Mundada
e-mail: monicamundada5@gmail.com

Santosh Gaikwad
e-mail: santosh.gaikwadcsit@gmail.com

Bharti Gawali
e-mail: bharti_rokade@yahoo.co.in

S.C. Satapathy et al. (eds.), *Proceedings of the International Congress on Information and Communication Technology*, Advances in Intelligent Systems and Computing 439, DOI 10.1007/978-981-10-0755-2_27

with the available phones and prosodic features from training speech database [1, 2]. The speech units are classified into phonemes, diaphones, and syllables. The output of speech synthesis system depends on the size of the speech units involved in the execution of the method. The designing of text-to-speech system is organized into two parts: front end and back end. In the front-end module, initially the input text comprising of symbols like numbers and abbreviations are transformed into the equivalent words. This process is defined as the text normalization, preprocessing, or tokenization. The text into prosodic units like phrases, clauses, and sentences are assigned with the phonetic transcriptions. The back-end phase produces the synthesis of the particular speech with the use of output provided from the front end. The symbolic representations from first step are converted into sound speech and the pitch contour, phoneme durations and prosody are incorporated into the synthesized speech.

The paper is structured in five sections. The techniques of speech synthesis are described in Sect. 2. Database for synthesis system is explained in Sect. 3. Section 4 explains speech quality measurement. Section 5 is dedicated with experimental analysis followed by conclusion.

2 Concatenate Synthesis

Concatenate speech synthesis is a method where speech is generated by concatenating speech units one after the other as per the requirement. There are three different types of concatenate speech synthesis. They are domain specific synthesis, diphone synthesis, and unit selection synthesis [2]. The focus of the paper is unit selection synthesis.

In this method, the database is built up with all phones present in the particular language. The design of such database includes well-labeled phones with high-quality utterances. The synthesized speech output signal is generated with the concatenated parts from the database [2]. The output speech produced in this method has greater impact on intonation, way of speaking style, and emotions associated with the speech. Also, the construction of large database corpus produces the high quality of synthesized speech. The USS (unit selection synthesis) extracts the prosodic and spectral part from input speech signal during the training part. In synthesis part, the analysis of text is done and prosody is incorporated with the use of algorithm and artificial speech is produced [2]. In USS, initially the text is converted into phones of the particular segment. Then the phones are assigned the labels like vowels, semivowels, and consonants. With the help of acoustic trees the IDs are generated for the given input. At the final step, the speech is synthesized with the USS algorithm with the incorporation of the needed prosody elements.

3 Hidden Markov Model-Based Speech Synthesis

Hidden Markov model synthesis is also called statistical parametric synthesis of speech. The significance of this method allows the variation in voice easily. In this method, the speech is synthesized on the basis of the parameters extracted from the recorded utterances. The HTS system, the context-dependent hidden Markov model generates the excitation parameters [3]. Thus they are treated as input for the speech waveforms in the later stage. In HMM-based speech synthesis there is no need for large database corpus and the quality of voice maintained [2]. In the two stages of HMM execution part initially the spectrum and excitation parameters are extracted from speech database and modeled by context-dependent hidden Markov model. In the next stage, context-dependent hidden Markov model is concatenated according to the text to be synthesized. Then spectrum and excitation parameters are generated from the hidden Markov model using a speech parameter generation algorithm. Both the techniques are implemented using Festival framework [4].

4 Speech Quality Measurement

Speech quality measurement is the level of audible and perceptible level of the output speech. There are two methods for performing this relative task.

4.1 Subjective Quality Measure

The quality associated with the produced speech depends on the paradigms of subjective perception and judgment process. In this method, the personal assessment is done from the individual with the help of mean opinion score (MOS) test. The subject evaluating the sentence assigns the grades depending on the quality of speech with respect to 5-point scale. In which the grade 1 is assigned to the least or unsatisfactory speech and the grade 5 to the excellent speech quality. In this the system is trained with the analysis performance report of each perceptual listener involved in the test [5, 6].

4.2 Objective Quality Measure

This method computes for the automated real-time quality measurement. The perceptual listener goes through the computational algorithm of the system. Real-time quality monitoring is gained only with the objective speech quality measurement. This method proves more reliable and accurate as compared to subjective listening experiments. In the objective quality measure mean square error (MSE) and peak signal-to-noise ratio (PSNR) techniques were used.

(a) Mean Square Error (MSE)

The mean squared error (MSE) is defined as the average of the squares of the errors, that is, the difference between the estimator value and the estimated value. The difference occurs because of randomness or because the estimator does not account for information that could produce a more accurate estimation of speech synthesis [7].

(b) Peak Signal-to-Noise Ratio (PSNR)

PSNR is measured in terms of the logarithmic decibel scale. The use of PSNR is to measure the quality of reconstruction of signal and image. The signal in this case is the original utterance, and the noise is the error introduced during the process of speech synthesis [8]. Peak signal-to-noise ratio (PSNR) is measured as the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the quality of its representation.

4.3 Signal-Based Quality Measure

In the signal-based quality measure the perceptual evaluation of speech quality (PESQ) technique [9–11] is followed. The current version P.862 of PESQ algorithm is highly reliable [12]. For this experiment we proposed MFCC features for the signal-based quality measure. Mel-frequency cepstral coefficients (MFCC) technique is robust and dynamic technique for speech feature extraction. The fundamental frequency, prosodic, energy variation in the syllable and many other features are studied with MFCC feature set. For the quality measure we extracted 13 features from synthesized speech and original speech file.

5 Speech Database

The speech database collected for this experiment includes the sentences from philosophy and short stories. The sentences were recorded by male and female speakers. Male speaker was with South Indian accent and female voice was with normal accent. The male and female both were from academic field and practiced the session. The recording was done in noise-free environment. The speech signal was sampled at 16 kHz. The set of 30 sentences were synthesized using unit selection and hidden Markov model. Noise-free lab environment with multimedia laptop speaker was used to play these utterances to the postgraduate students. The students were of age group 22–25, with no speech synthesis experience.

6 Experimental Analysis

Analysis of Mean Opinion Score (MOS)

MOS is calculated for subjective quality measurement. It is calculated for the synthesized speech using the unit selection synthesis and HMM approach. It was counseled to the listeners that they have to score between 01 and 05 (Excellent—05; Very good—04; Good—03; Satisfactory—02; Not understandable—01) for understandability. The mean of the scores given by each individual subject for ten sentences of the unit selection approach is shown in Table 1. The details of MOS score obtained from HMM speech synthesis method for ten sentences are shown in Table 2.

Table 1 Unit selection speech synthesis of the scores given by each subject for each synthesis system

Subject										
Sentence	1	2	3	4	5	6	7	8	9	10
1	5	5	5	5	4	4	5	4	4	5
2	5	5	4	5	5	4	5	4	4	5
3	4	4	5	4	3	3	4	2	5	4
4	5	4	4	5	4	4	5	5	5	5
5	5	5	5	5	4	4	5	3	3	5
6	5	4	5	5	5	4	5	4	4	5
7	4	5	4	4	4	4	4	4	4	4
8	4	4	5	4	4	5	4	5	5	4
9	5	3	5	5	3	4	5	3	5	5
10	5	5	4	5	4	4	5	4	4	5

Table 2 HMM-based speech synthesis of the scores given by each subject for each synthesis system

Subject										
Sentence	1	2	3	4	5	6	7	8	9	10
1	4	5	5	5	5	4	5	4	5	3
2	3	5	4	4	3	4	3	3	3	4
3	5	4	4	4	4	3	4	4	3	5
4	5	4	4	4	3	4	4	5	4	3
5	3	4	5	5	5	3	4	4	5	4
6	2	4	4	3	4	2	4	5	4	4
7	3	5	5	2	5	1	5	3	3	5
8	4	4	4	1	4	2	5	4	2	3
9	4	3	3	3	5	3	4	5	2	4
10	5	4	4	1	2	2	4	4	4	5

Table 3 Mean and variance of the scores obtained across the subjects from unit selection and HMM approach

Subject	Unit selection method		HMM synthesis approach	
	Mean score	Variance	Mean score	Variance
1	4.56	0.25	3.90	1.19
2	4.23	0.52	2.73	1.71
3	4.03	0.79	2.56	1.35
4	4.56	0.25	2.80	1.61
5	4.10	0.43	2.46	0.947
6	4.03	0.37	3.10	1.05
7	4.56	0.25	2.80	1.68
8	3.96	0.72	2.33	1.26
9	4.16	0.62	2.73	1.37
10	4.63	0.24	2.63	1.48

The mean and variance of the score obtained according to the subject using the two experimental approaches, i.e., unit selection and HMM-based speech synthesis approach are shown in Table 3.

It is observed that from Tables 3 and 4 mean scores increase with the increase in the syllable coverage.

(a) PSNR and MSE Quality Measure

The PSNR and MSE methods were used for subjective quality measure of speech synthesis based on hidden Markov model and unit selection approach. Table 4 represents the MSE and PSNR values for unit selection-based speech synthesis. HMM-based speech synthesis using MSSE and PSNR is shown in Table 5.

Table 4 MSE and PSNR values for unit selection-based speech synthesis

Sr. No	Original speech file	Synthesized file	MSE	PSNR
1	A1	a1	7.94	3.30
2	A2	a2	4.57	6.72
3	A3	a3	1.02	3.21
4	A4	a4	3.70	4.20
5	A5	a5	7.61	2.57
6	A6	a6	5.32	1.26
7	A7	a7	8.06	7.56
8	A8	a8	7.20	1.29
9	A9	a9	9.25	3.24
10	A10	a10	7.01	4.08
Average			6.168	3.743
Quality (100-Average)			93.83	96.26

Table 5 MSE and PSNR values for hidden Markov model speech synthesis

Sr. No	Original speech file	Synthesized file	MSE	PSNR
1	B1	b1	9.15	7.315
2	B2	b2	8.38	6.24
3	B3	b3	13.5	5.25
4	B4	b4	10.4	8.40
5	B5	b5	9.26	8.76
6	B6	b6	9.38	9.42
7	B7	b7	10.10	9.05
8	B8	b8	9.63	6.56
9	B9	b9	10.42	8.49
10	B10	b10	12.40	7.44
Average			10.26	7.69
Quality (100-Average)			89.73	92.31

Table 6 Comparative result of unit and HMM speech synthesis

Sr. No	Approach of synthesis	MFCC mean (%)	MFCC STD (%)	MFCC var (%)	MSE (%)	PSNR (%)
1	HMM	80	80	80	89.73	92.31
2	Unit selection	90	90	80	93.83	96.26

Table 6 shows the comparative performance of both unit and HMM for accent recognition using MFCC, MSE, and PSNR techniques.

From Table 6, it is observed that the unit selection-based accent identification gives better performance than HMM-based speech synthesis.

(b) Mel-frequency Cepstral Coefficients

The signal-based synthesis quality measure is experimented for unit selection and hidden Markov model-based speech synthesis. For the performance variation, we calculated the mean, standard deviation, and variance as a statistical measure. The details of sentences and label used for the unit selection-based speech synthesis are described in Table 7.

The MFCC mean-based performance of unit selection-based synthesis is shown in Table 8. Table 9 represents the details of standard deviation of MFCC for unit selection speech synthesis.

The sentences used for hidden Markov model-based synthesis using MFCC-based method is shown in Table 10. The details of performance of MFCC mean and standard deviation for hidden Markov model-based speech synthesis are shown in Tables 11 and 12, respectively.

Table 7 Sentences and label used for unit selection-based speech synthesis

Sr. No	The original sentence	Label used for original speech file	Label used for synthesis speech file
1	Will we ever forget it	A1	a1
2	There was a change now	A2	a2
3	I had faith in them	A3	a3
4	She turned in at the hotel	A4	a4
5	We'll have to watch our chances	A5	a5
6	It was a curious coincidence	A6	a6
7	There was nothing on the rock	A7	a7
8	I have no idea, replied Philip	A8	a8
9	Anyway, no one saw her like that	A9	a9
10	Surely I will excuse you, she cried	A10	a10

Table 8 The performance of MFCC mean-based unit selection speech synthesis

Synthesized speech											
Original speech signal		**a1**	**a2**	**a3**	**a4**	**a5**	**a6**	**a7**	**a8**	**a9**	**a10**
	A1	**0.120**	3.242	2.31	1.392	3.42	3.008	2.983	5.094	7.01	1.234
	A2	1.281	**0.009**	2.111	2.453	7.632	1.90	4.02	1.223	8.01	3.04
	A3	1.453	3.21	**0.080**	3.25	1.99	2.843	3.921	2.963	2.093	6.70
	A4	3.02	1.230	2.564	**0.899**	2.786	5.453	1.672	1.981	2.67	3.45
	A5	2.40	1.450	5.432	2.932	**0.021**	3.921	6.05	4.675	7.00	3.674
	A6	3.896	8.09	3.983	2.732	3.674	**0.673**	2.843	5.03	3.894	4.92
	A7	1.893	1.563	2.03	2.100	4.92	3.67	**1.460**	1.273	3.521	7.38
	A8	2.932	3.674	2.732	3.721	3.567	2.732	3.876	**0.783**	2.673	3.643
	A9	1.776	2.732	4.332	5.893	2.783	3.874	2.743	4.87	**1.091**	3.021
	A10	2.873	2.983	1.873	1.90	2.763	1.563	4.02	1.788	2.032	4.328

Table 9 The performance of MFCC STD-based unit selection speech synthesis

Synthesized speech											
Original speech signal		**a1**	**a2**	**a3**	**a4**	**a5**	**a6**	**a7**	**a8**	**a9**	**a10**
	A1	**0.456**	1.200	1.892	3.902	3.872	2.893	5.783	3.872	4.500	2.673
	A2	1.231	**0.632**	2.762	2.090	1.988	2.763	1.235	1.6532	4.673	6.011
	A3	5.632	3.982	**1.050**	1.928	2.782	2.782	1.892	1.292	3.020	5.873
	A4	3.092	2.093	4.092	**1.837**	4.932	3.091	5.781	3.982	2.983	2.983
	A5	1.882	1.0291	3.0281	3.091	2.872	**1.092**	4.092	3.982	3.982	5.021
	A6	3.091	4.873	3.982	2.983	1.022	**0.932**	1.829	2.893	4.092	4.093
	A7	2.983	3.092	2.993	4.984	2.831	8.011	**1.920**	1.892	3.001	3.092
	A8	5.011	3.921	4.984	5.092	2.931	4.982	1.092	**0.982**	1.778	4.832
	A9	2.938	5.011	2.932	6.091	2.983	1.921	2.932	4.632	**1.092**	1.920
	A10	3.092	1.821	1.9082	3.921	4.921	3.842	4.983	4.530	2.321	**0.210**

Table 10 Sentences and label used hidden Markov model speech synthesis

Sr. No	The original sentence	Label used for original speech file	Label used for synthesized speech file
1	Gad, do I remember it	B1	b1
2	I can see that knife now	B2	b2
3	They robbed me a few years later	B3	b3
4	Now, you understand	B4	b4
5	He caught himself with a jerk	B5	b5
6	How does your wager look now	B6	b6
7	It won’t be for sale	B7	b7
8	Now it was missing from the wall	B8	b8
9	It is the nearest refuge	B9	b9
10	He can care for himself	B10	b10

Table 11 The performance of MFCC mean-based hidden Markov model speech synthesis

Synthesized speech											
Original speech signal		**b1**	**b2**	**b3**	**b4**	**b5**	**b6**	**b7**	**b8**	**b9**	**b10**
	B1	**0.234**	3.781	5.155	6.280	2.662	5.442	3.601	5.432	3.970	11.950
	B2	5.227	**0.200**	8.700	6.191	1.7100	5.327	5.465	8.932	2.242	6.126
	B3	1.900	3.815	**0.210**	9.044	3.123	1.090	2.120	3.030	5.445	9.580
	B4	5.559	0.934	1.980	**0.936**	1.2315	1.780	2.090	2.050	6.318	12.272
	B5	2.980	3.800	3.178	2.153	**0.119**	2.130	2.150	3.092	2.339	23.011
	B6	2.051	9.1400	**0.221**	1.050	1.781	3.873	1.363	1.030	3.335	16.09
	B7	2.463	4.990	5.900	2.191	2.130	2.172	**0.181**	1.192	2.991	8.700
	B8	4.839	6.566	2.550	2.781	1.630	1.152	0.800	**0.500**	5.344	9.811
	B9	3.992	7.502	3.300	2.050	1.980	1.050	1.262	2.111	**0.300**	7.020
	B10	7.793	6.176	3.528	2.128	6.512	7.900	3.512	1.393	**0.810**	1.23

Table 12 The performance of MFCC STD-based hidden Markov model speech synthesis

Synthesized speech											
Original speech signal		**b1**	**b2**	**b3**	**b4**	**b5**	**b6**	**b7**	**b8**	**b9**	**b10**
	B1	**0.110**	1.221	2.119	1.290	1.890	1.233	2.030	3.01	1.178	5.020
	B2	2.120	**1.20**	1.900	3.402	7.680	8.900	11.02	2.678	2.343	7.890
	B3	0.900	1.123	**2.342**	3.784	3.134	4.030	2.178	9.01	2.05	5.030
	B4	2.564	3.100	2.870	**0.900**	1.760	2.345	2.403	3.050	2.870	3.435
	B5	1.890	1.450	2.123	**0.200**	5.40	2.656	1.934	2.999	7.030	9.01
	B6	0.890	1.543	2.212	3.210	3.521	**0.321**	2.986	1.776	2.832	3.02
	B7	2.320	1.564	2.220	5.040	3.442	5.021	**0.210**	3.450	3.887	7.00
	B8	1.747	2.030	5.020	2.456	3.022	**0.884**	3.007	4.302	2.345	1.284
	B9	2.336	3.020	3.040	3.121	3.998	7.060	4.990	3.2020	**1.02**	1.998
	B10	1.987	2.030	2.0440	2.312	3.220	1.228	6.070	3.009	4.006	**0.320**

7 Conclusion

The quality of speech synthesis is experimented using MOS score, MSE, PSNR, MFCC-based techniques for hidden Markov model (HMM) and unit selection synthesis (USS) approach. The MFCC-based method is evaluated using the mean, standard deviation, and variance. For all the estimated methods the unit selection method gives a better performance than hidden Markov model techniques as the database is small.

References

1. Mohammed Waseem, C.N Sujatha, “Speech Synthesis System for Indian Accent using Festvox”, International Journal of Scientific Engineering and Technology Research, ISSN 2319-8885 Vol. 03, Issue. 34 November-2014, Pages: 6903–6911.
2. Sangramsing Kayte, Kavita Waghmare, Dr. Bharti Gawali “Marathi Speech Synthesis: A review” International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 6, 3708–3711.
3. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis” In Proc. of ICASSP 2000, vol 3, pp. 1315–1318, June 2000.
4. A. Black, P. Taylor, and R. Caley, “The Festival Speech Synthesis System. System documentation Edition 1.4, for Festival Version 1.4.3 27th December 2002.
5. Series P: Telephone Transmission Quality “Methods for objective and subjective assessment of quality” Methods for Subjective Determination of Transmission Quality ITU-T Recommendation P.800.
6. ITU-T P.830, Subjective performance assessment of telephone-band and wideband digital codecs.
7. Lehmann, E. L.; Casella, George. “Theory of Point Estimation (2nd ed.). New York: Springer. ISBN 0-387-98502-6. MR 1639875.
8. Huynh-Thu, Q.; Ghanbari, M. (2008). “Scope of validity of PSNR in image/video quality assessment”. Electronics Letters 44 (13): 800. doi:10.1049/el:20080522.
9. SR Quackenbush, TP Barnwell, MA Clements, Objective Measures of Speech Quality (Prentice-Hall, New York, NY, USA, 1988).
10. AW Rix, MP Hollier, AP Hekstra, JG Beerends, PESQ, the new ITU standard for objective measurement of perceived speech quality—part 1: time alignment. Journal of the Audio Engineering Society 50, 755–764 (2002).
11. JG Beerends, AP Hekstra, AW Rix, MP Hollier, PESQ, the new ITU standard for objective measurement of perceived speech quality—part II: perceptual model. Journal of the Audio Engineering Society 50, 765–778 (2002).
12. ITU-T P.862, Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech. codecs 2001.