# Toward Efficient Low Cost Highly Accurate Emotion speech Synthesizer

**Ahmed Mustafa, Aisha-Hassan A. Hashim, Othman Khalifa & Shihab Ahmed**
Faculty of Engineering, International Islamic University Malaysia

**Abstract:**
A Text to Speech (TTS) system with the ability to express emotions is an interesting technology that is still under development. There have been multiple proposals to simulate emotion so far, and there are multiple dimensions for assessment. No system guarantees high score in all of these dimensions, this means that no system works in a direction to get low computation load, small database along with high accuracy and excellent voice quality. After all of these qualities are relative and fuzzy and there is no rigid grading system. In this paper we will propose a new path for research that will work toward improving all of the quality factors together, so that future work can come up with a more optimum solution for the emotional TTS systems.

## Introduction:

Text to speech system capable to express emotional state along with speech is an undergoing work of research, as there has been no standardized solution for this issue (1). Multiple proposals for emotional simulation have been suggested in previous research, depending on the Speech synthesis approach. Two primary technologies for speech synthesis are being followed which are concatenative synthesis and formant synthesis (2). Formant synthesis which is the old trend tends to generate sound sample from scratch by simulating the actual mouth speech mechanism. This is will achieve the lowest quality of voice and high computational load, but extreme low or zero databases sizes. Concatenative is the modern approach and mostly achieves the best quality, it includes two main approaches: Unit selection and diaphone synthesis. Unit selection has proved superior quality over diaphone (3) (4), but it has its main drawback which is complex processing algorithms and large databases (5) (6). Diaphone analysis on the other hand has the advantage of small database needed, had been constantly improving its quality thanks to the DSP algorithm known as PSOLA (Pitch Synchronous Overlap Add Method). The MBR-PSOLA algorithm (7) puts diaphone TTS speech quality in a position comparable to that of the unit selection TTS system (8).

Emotional synthesis using efficient and light weight methods such as diaphone TTS have had its share of the research work. However after adding emotion to the TTS, the results have been achieved so far shows a considerable flaw in the degree of accuracy and voice quality, to the point that disqualifies it from the industry standards (9). Diaphone based emotional system had accuracy values as low as 44.2% which was recorded for recognition rate for its Happy expression capability, and a quality as low as 2.7 out of 5 (10). Thus there is an obvious problem in the issue of emotion quality in diaphone based systems. If we manage to improve the speech quality in diaphone systems, we will be working toward the goal of combing the low size and cost of the diaphone system with this accurate emotional synthesis.

Previous research (9) (11) (12) related to emotion in diaphonic TTS argued that there are few acoustic parameters that needs to be considered in order to achieve emotional expression, such as F0 mean, F0 Range , segmental modification of duration, pitch, energy and spectral envelope parameters for diaphones. They tried to find the correlation between these parameters and emotional expression through their signal analysis. Our view is that these parameters are the effects of some other hidden causes that need to be considered in order to achieve high accuracy results, i.e. they are not the cause of emotional expression.

In this paper we will present our proposed approach to extract the hidden factors highlighted above, and implement a speech synthesis to demonstrate the effect of considering these factors in speech synthesis, using the TD-PSOLA for backend signal processing. We will then verify our results using standardized perceptual experiments used for evaluation of previous research (11) (2) (12) (5). In this type of experiments as many as 10 naïve listeners will evaluate the accuracy of simulation by recording their perception on the tested samples. Provided that the spoken meaning of those sample has no relation to the emotional speech and they have no knowledge of which emotional state we are trying simulate. We will follow the guidelines imposed

Manuscript received September 5, 2008.
Manuscript revised September 20, 2008.

by previous research (13) (14) (15)related to evaluation methodology in order to achieve valid test data. To further validate the test we conducted the same evaluation test on the input analysis sample.

In the next section we will explain the analysis that we have done in order to come up with our hypothesis, section 3 will show the implementation of this hypothesis, and section 4 will explain the details of our perceptual evaluation experiment. Section 5 will show analysis to the result data of the evaluation and section 6 will present the conclusion of this research.

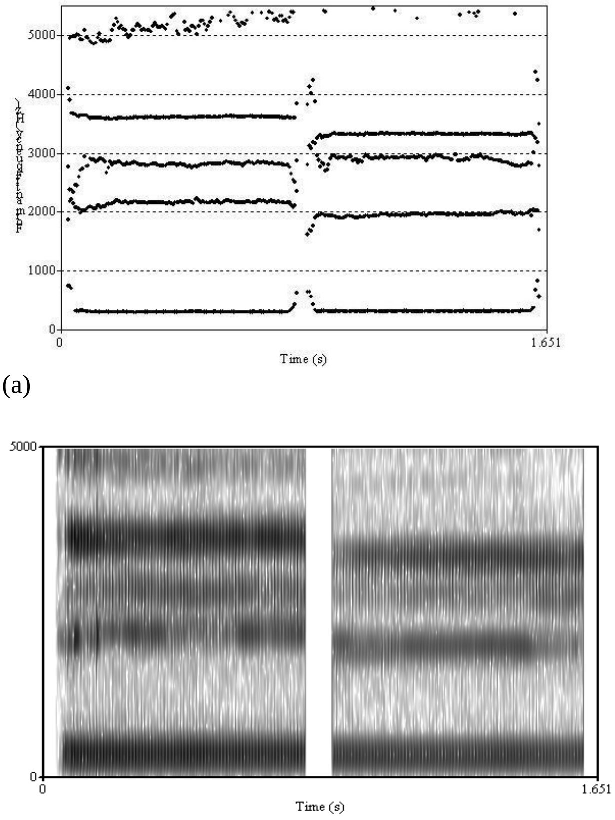
## Spectral analysis:

Before we can begin explaining the hypothesis we have to explain some background about speech mechanics.

Human mouth is basically a speech system that is composed of 2 main components: a signal generator and a complex filter. The vocal cords produce a signal similar to a pulse train which has a frequency spectrum that is almost a flat rich envelope that spans wide range of frequencies. This signal is passed to the oral cavity which acts as a complex low pass filter that adds harmonics to the signal and filter parts of its spectrum to generate the phoneme sound. The shape of the mouth and the articulation operation taking place will determine the spoken letter or phoneme, and the shape of the frequency spectrum envelope.

The main motivation behind our research comes from the common scene fact that when people are happy they smile, and when the talk happily they talk smiling. Form this starting point we can say that the shape of the oral cavity varies with emotional state. This means that the filter applied to generate the phoneme spoken under happy emotion is in fact different from the filter applied to generate the same phoneme under neutral state.

To put this in accountable terms, we have recorded two sample of the same word under happy and neutral emotion, and fixed the parameters manipulated by previous researchers i.e. pitch, mean F0, duration in both of them. We have done formant analysis on the two recorded sample and found out that these two samples actually have deferent formant structure. Thus the two samples which have similar acoustic parameters express different emotion because they have different spectral formant. And this lead to the hypothesis that achieving different emotion in TTS need manipulating the formant of the diaphones, to shift or add new harmonics.



(a)

(b)

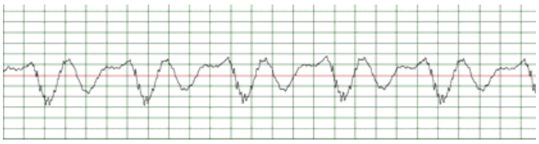
**Figure 1** shows the difference between formants, in (a) we can see the harmonic structure of the neutral sample of the phoneme /a/ and the right is the same phoneme under happy emotion. Note that the first harmonic F0 is the same which proves there is no pitch difference. (b) shows the spectrogram of these 2 samples.

Achieving this formant manipulation requires spectral change which is not an easy operation, and infact it is almost impossible to add harmonics to a signal and in the same time preserve its natural feeling. This explains the hardship experienced by previous research which neglected this feature while trying to achieve the emotional expression. Unit selection didn’t face this problem as in unit selection full chunks of spoken words in which spectrum is already modified by the speaker’s mouth under a certain emotional state are used and there is no mean of modifying single neutral mono-phones or diaphones to achieve emotion.

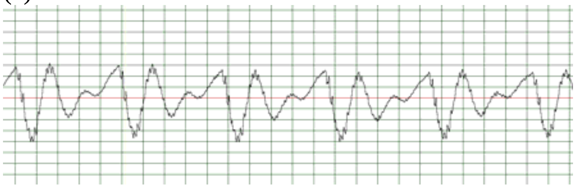
In order to boost the accuracy of diaphone synthesis we propose distinguishing phonemes under different emotional state as two separate diaphones, and using a new algorithm for diaphone concatenation that can accommodate emotional diaphones, to preserve the quality. If we manage to manipulate pitch and duration through this modified concatenation in a natural way, we can guarantee very high accuracy results for emotion as it has been recorded from natural voice under that target emotion.

## Expressive Diaphone concatenation:

Implementing the concatenation for emotional diaphone is somehow a straight forward operation. Based on previous research in this field (8) we will choose the TD-PSOLA algorithm. However in order to compress the database size we will encode the diaphones using the Multiband excitation model used by MBR-POLA in the preprocess stage. We will not go all the way with the MBROLA algorithm because it has its own drawbacks (16) on quality which is our main consideration. Thus we will implement this system and we will also normalize the samples in the preprocess stage and modify the pitch to achieve constant period $T_0$ to eliminate the pitch marker files, but we will not regenerate the data, to preserve the original harmonic pitch and by pass the re-synthesis stage in the MBROLA.



(a)



(b)

**Figure 2**: (a) shows the raw recorded data(b) shows the processed waves

## Experiment:

In order to evaluate the accuracy and quality of this system we will have to conduct perceptual experiment on 10 naïve listeners using multiple samples. The experiment will be conducted to prove the fact the formant shape has the biggest share of emotional expression and it is superior on other parameters such as mean pitch, duration, and F0 variance. To do this the sample synthesized using our system will not poses any pitch variation and will keep static acoustic parameters such as energy variation. The result of the perceptual survey given to the listeners will show how significant or algorithm is.

In order to validate this testing mechanism, we will feed the analysis data which has been generated by professional speakers to the listeners to see if their evolution is correct with this natural recorded samples. We have also conducted validity and reliability test for our evaluation mechanism. The correlation with the perfect set of data on the last row will show how valid is our testing mechanism.

| | Joy | Neutral | Fear | Restless | Anger | Surprise | affection | Secure | Disgust | Bored |
|---|---|---|---|---|---|---|---|---|---|---|
| Joy | 94 | 7 | 0 | 2 | 0 | 5 | 6 | 1 | 0 | 1 |
| | 81.03 | 6.03 | 0.00 | 1.72 | 0.00 | 4.31 | 5.17 | 0.86 | 0.00 | 0.86 |

Table 1

Correlation: 0.995901 = 99.5%

## Evaluation:

| | Joy | Neutral | Fear | Restless | Anger | surprise | affection | Secure | Disgust | Bored |
|---|---|---|---|---|---|---|---|---|---|---|
| Joy | 85 | 16 | 0 | 2 | 0 | 5 | 5 | 2 | 0 | 1 |
| % | 73.28 | 13.79 | 0.00 | 1.72 | 0.00 | 4.31 | 4.31 | 1.72 | 0.00 | 0.86 |
| Neutral | 3 | 98 | 0 | 1 | 0 | 6 | 3 | 4 | 0 | 2 |
| % | 2.59 | 84.48 | 0.00 | 0.86 | 0.00 | 5.17 | 2.59 | 3.45 | 0.00 | 1.72 |

Table 2

Table 1 shows that, the correlation value shows that the test has a high validity above the significance level (40%).

After validating the perceptual testing mechanism, we will generate 2 sets of 116 sentences with the same content under neutral and happy emotion. We will make an evaluation run on these results and analyze it.

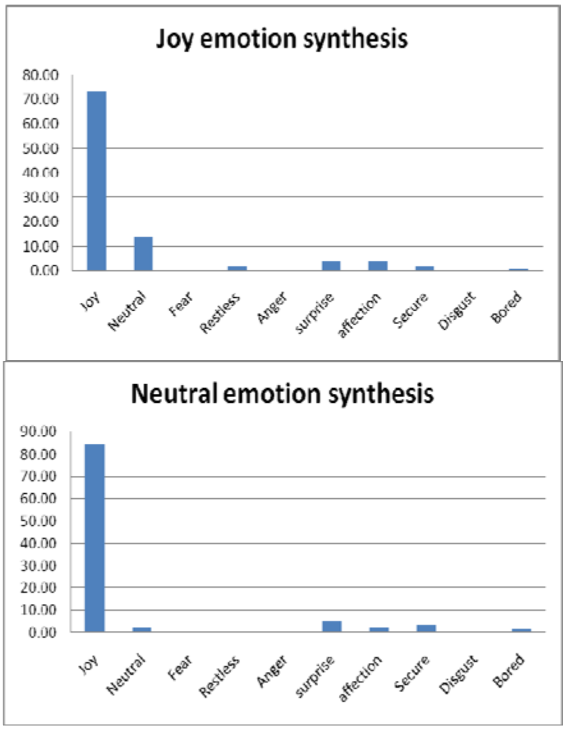


**Figure3:** the plot of the experiment result data for both neutral and happy emotion

Table 2 shows the results obtained from the experiment highlighted above. Each row present one set of test data. We can see that our system have synthesized the neutral and the joyful samples using now pitch or energy variation and only depending on formant variation. The accuracy is high enough to show that the formant based algorithm is significant enough to simulate emotion.

## Conclusion:

In this paper we have studied the effect of formant changes on the expressiveness of emotion, and we have assumed that it has a superior effect on the process of emotion generation in TTS using diaphone method. We have verified our hypothesis on happy emotion since it was the hardest emotion to simulate and we have achieved considerable improvement over previous methods that used diaphone synthesis. We have paved the way for future research that tends t utilizes the diaphone technology to simulate emotional expression with high accuracy. We have provided a solution to the problem of efficient low cost, highly accurate TTS system.

## Future work:

Future work is needed to implement this system on other emotional states such as anger and fear, and also to enhance the voice quality by combining pitch modification with the emotional diaphones to achieve in order to combine accuracy with quality.

## References

[1] Verification of acoustical correlates of emotional speech using formant synthesis. Burkhardt, F., & Sendlmeier, W.F. Northern Ireland. : Proceedings of the ISCA Workshop on Speech and Emotion, 2000. 151-156.

[2] Data-driven formant synthesis. Carlson, R., Sigvardson, T., & Sjölander, A. Stockholm, Sweden: KTH : Proceedings of Fonetik, 2002 - speech.kth.se, 2002.

[3] Perfect synthesis for all of the people all of the time. Black, A.W. 11-13 Sept. 2002 Page(s): 167 - 170, s.l. : Proceedings of 2002 IEEE Workshop on, 2002. 10.1109/WSS.2002.1224400.

[4] Unit selection and emotional speech. Black, A.W. Geneva, Switzerland : Proc. Eurospeech 2003, 2003.

[5] Unit selection in a concatenative speech synthesis system using alarge speech database. Hunt, A.J. Black, A.W. ATR Interpreting Telecommun. Res. Labs., Kyoto. 1 pages: 373-376, s.l. : Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 7-10 May 1996, Vol. I. 0-7803-3192-3.

[6] Informed Blending of Databases for Emotional Speech Synthesis. Gregor O. Hofer, Korin Richmond, Robert A.J. Clark. s.l. : In Proc. Interspeech, September 2005.

[7] MBR-PSOLA: Text to Speech synthesis based on a MBE re-synthesis of the segments database. T. Dutoit, H. Leich. s.l. : Speech Communication 13, 1993.

[8] "A comparison of Four candidate Algorithms in the context of High Quality Text to Speech Synthesis. T. Dutoit, H. Leich. s.l. : ICASSP'94, 1994.

[9] Prosodic analysis of a multi-style corpus in the perspective of emotional speech synthesis. Enrico Zovato, Stefano Sandri, Silvia Quazza, and Leonardo Badino. s.l. : Proc. ICSLP 2004, Jeju, Korea, 2004.

[10] Expressive speech synthesis using a concatenative synthesizer. Murtaza Bulut, Shrikanth S. Narayanan, Ann K. Syrdal. Denver, Colorado : ICSLP, , September 2002.

[11] Investigating the role of phoneme-level modifications in emotional speech resynthesis. C. Busso, S. Yildirim, A. Kazemzadeh, C. M. Lee, S. Lee, S. Narayanan. Lisbon, Portugal : Eurospeech, Interspeech,, September 2005.

[12] Towards emotional speech synthesis: a rule based approach. Enrico Zovato, Alberto Pacchiotti, Silvia Quazza, Stefano Sandri. Carnegie Mellon University Pittsburgh the City : 5th ISCA Speech Synthesis Workshop, 14th-16th June 2004.

[13] Vocal communication of emotion: A review of research paradigms. Scherer, K.R. 40, s.l. : Speech Communication,, 2003, Vols. 1-2. 227-256.

[14] Of all Things the Measure is Man. Automatic Classification of Emotions and Inter-labeller Consistency. Steidl, S., Levit, M., Batliner, A, Nth, E, Niemann, H.,. s.l. : ICASSP, 2005.

[15] Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation. Campbell, N.,. s.l. : ICSLP, 2004.

[16] Improving quality of MBROLA synthesis for non-uniform units synthesis. Bozkurt, B., et al. page 7-10, s.l. : Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop, 11-13 Sept. 2002 .