Arabic Unit Selection Emotional Speech Synthesis using Blending Data Approach

Waleed M. Azmy
Faculty of computers and information
Cairo University
Cairo - Egypt

Sherif Abdou
Faculty of computers and information
Cairo University
Cairo - Egypt

Mahmoud Shoman
Faculty of computers and information
Cairo University
Cairo - Egypt

ABSTRACT
This paper introduce the work done to build an Arabic unit selection voice that could carry emotional information. Three emotional sates were covered; normal, sad and questions. An emotional speech classifier was used to enhance the intelligibility of the used recorded speech database. The classification information was employed in the proposed target cost to produce more natural and emotive synthetic speech. The system is evaluated according to the naturalness and emotiveness of the produced speech. The system evaluations show significant increase in the naturalness and emotiveness scores.

Keywords
Speech synthesis, Emotion, Festival, Emovoice.

1. INTRODUCTION
Speech synthesis technologies have been improved over the last decades such that the output of the synthesized speech approaches some naturalness of the prerecorded data. In order to increase the acceptability and improve the performance of the synthesized speech the syntheses of attitude and emotion have to be considered [1]. Generating emotional speech can be applied on various applications like generating sales dialogues, generating social interactions in an interactive soap opera setting [5] and in gaming. A more abstract feasibility study is the combination of speech synthesis with a photo-realistic facial animation model [7].

Attempts to add emotion effects to synthesized speech have existed for more than a decade. Several prototypes and fully operational systems have been built based on different synthesis techniques, and relatively quite number of smaller studies has been conducted [2]. Approaches towards producing emotionally synthetic speech have changed considerably over years. Early systems, including formant and diphone systems, have been focused around "explicit control" models; early unit selection systems have adopted a "playback" approach [4]. Concatenative unit selection speech synthesis systems are generally found to produce speech that sounds more similar to the target speaker than statistical parametric speech synthesis systems do [3]. So, in our system we adopted the unit selection approach to build an emotional TTS voice. Three emotional states which are normal, sad and question have been used. The Festival Speech Synthesis System [8] was used as a basis for building this voice. It was slightly modified to fit the requirements for an emotional speech synthesizer.

The Arabic language is one of the widely spoken languages in the word with around 480 million people speaks Arabic as their mother language. Arabic diphone based voices were developed using n the Festival system [15][16]. Also high quality concatenative based Arabic voices were developed for commercial systems [17]. To our knowledge there are no reported attempts for developing emotional expressive Arabic voices..In this work we developed an emotional Arabic TTS voice. The database used for building that voice is RDI Arabic Saudi male database. This database consists of 10 hours of recorded speech with neutral emotion. It also includes one hour of recordings for the same speaker with four different emotions which are sad, happy, questioning and surprise. Since the emotional part of this database is clearly small we increased its size using the Blending data approach. An Emotional classifier is trained on the emotional part of the database and is used to detect any recorded phrases with blended emotion in the whole database. Also a new emotional target cost is proposed to employ the classification information in the synthesis process.

This paper is organized as follows: in Section 2 the process of building emotional unit selection synthetic voices is described, with details of the database types and the Multisyn, unit selection approach in festival system. Section 3 describes the Arabic emotional speech database used in this work. Section 4 introduces the Emovoice; the emotional speech classifier and the proposed setup to be used in creating emotional database for emotional synthesis systems. Section 5 gives the details of the proposed target cost algorithm. Section 6 describes the design of the evaluation process, including the evaluation metrics, scenarios and results. Section 7 concludes the main findings and final conclusions.

2. EMOTIONAL UNIT SELECTION

2.1 Unit Selection syntheses with emotions
Unit Selection Synthesis, where appropriate units are selected from large databases of natural speech, has greatly improved the quality of speech synthesis [9]. But the quality improvement has come at a cost. The quality of the synthesis relies on the fact that little or no signal processing is done on the selected units, thus the style of the recording is maintained in the quality of the synthesis. The synthesis style is implicitly the style of the database. If it is required more general flexibility, more data of the desired style have to be recorded. Unit selection techniques will provide synthesizers with the quality of the database they are built from. Thus it is possible to synthesize various emotions if there exists a recorded database of the appropriate type

Unit selection uses a cost function to select segments that vary in length from a large database of speech. The selected segments or units are concatenated to form the speech signal. Signal processing after the concatenation of units is usually not conducted because of artifacts. This is also one of the biggest drawbacks of unit selection: most unit selection systems do not have an implementation that allows for modification of any acoustic parameters, therefore making it non-trivial to use in emotional speech synthesis.

Two measures are used to find the best units. The join cost that gives an estimate of how well two units join together and the target cost that describes how well a candidate unit match the target unit. Instead of having a separate voice for each emotion, in our system we introduced an emotional target cost that can be added to the target cost function to find the most suitable units for synthesizing speech with the desired target emotion.

2.2 Emotional speech Database Structure

One can construct different voices for different tasks (e.g. weather, stocks, and email reading) and make explicit changes in voice when changing domains. It is called tiering. The second approach is combining domain related prompts together into a single voice, typically with a significant amount of prompts to support general synthesis. This is called blending [9] [10].

Tiering technique can work well if there are well defined borders between the voice types and the voice is not too large. Limited domain synthesis, where each style is for a particular domain, might be a good application for this technique. For example, one domain is to tell good news where the other domain tells bad news. Naturally, creating a large number of databases for different emotions will take a long time and is not trivial. It might be sufficient and more efficient to have a general database and just a few important units recorded in certain styles.

Blending allows, potentially, a smaller footprint and also less firm boundaries between the domains, thus switching between voice types is not required. Blended voices are harder to get.]. But blending technique works well in mixed domain based synthesis with other domain based databases and/or general ones, though it helps if they are basically in the same style. Mixing domain-based and general databases in a blended voice can produce excellent quality.

2.3 MutiSyn: Unit Selection Module in Festival

Since Festival is primarily a research toolkit, it provides a state-of-the-art unit selection synthesis module called Multisyn that satisfies two design goals. The first goal is to provide a stable general purpose unit selection implementation that is suitable for carrying out further research into unit selection and related techniques. The second goal is to provide the end user with a simple, mostly automatic mechanism to build their own voice for the system [11].

2.3.1 Utterance Structure

Festival uses a data structure called an utterance structure that consists of items and relations [12]. The utterance structure lies at the heart of Festival. An utterance represents some chunk of text that is to be rendered as speech. Default Festival architecture come with a pre-defined set of items linked with relations. The most important relation are [12]:

- *Token*: a list of trees. a list of tokens found in a character text string. Each root's daughters are the word's that the token is related to.
- *Word*: a list of words. These items also appear as daughters (leaf nodes) of the Token relation. They are also the leafs of the Phrase relation.
- *Phrase*: a list of trees. This is a list of phrase roots whose daughters are the words within those phrases.
- *SylStructure*: a list of trees. Each Word is the root of a tree whose immediate daughters are its syllables and their daughters in turn as its segments.
- *Syllable*: a list of syllables. In that relation its parent will be the word it is in and its daughters will be the segments that are in it.
- *Segment*: a list of segments (phones). Each member will be leaf nodes in the SylStructure relation. These may also be in the Target relation.

Figure 1 shows an example utterance structure. This example shows the word relation and the syntax relation. The syntax relation is a tree and the word relation is a list. The links are connecting the black dots. The items which contain the information are shown in the rounded boxes.

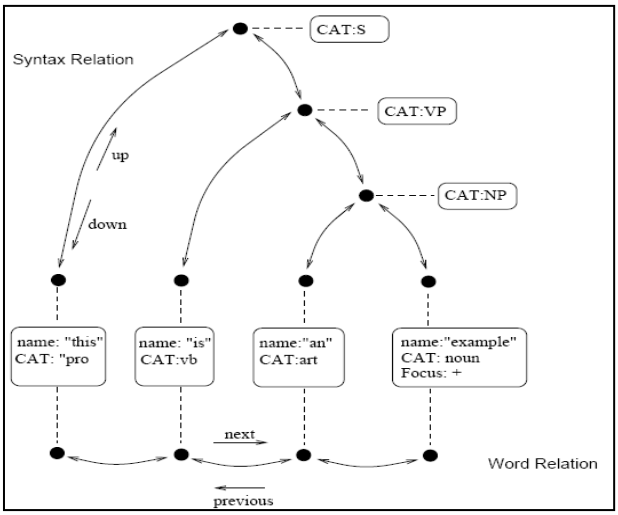


Figure 1. Utterance Structure Example

Each Item in the utterance structure has some features. Utterance can be modified to accept new Items and new features to be accessible through the synthesis process.

3. ARABIC SPEECH DATABASE

3.1 Database Description

At the core of any good unit selection speech synthesizer is the voice database. For our system we used RDI TTS Saudi speaker database. This database consists of 10 hours of recording with neutral emotion and one hour of recordings for four different emotions that are sadness, happiness, surprise and questioning. All databases are recorded from a male professional radio announcer at sample rate 16 kHz. The EGG signal is recorded with each utterance to support pitch synchronous analysis and pitch marking if necessary during the synthesis process. HMMs based Viterbi alignment procedure is used to produce the phone level segmentation boundaries of this database.

3.2 Voice Building for Multisyn

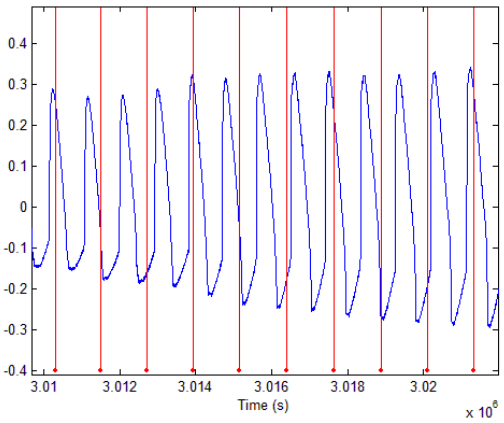
For building a unit selection voice in festival Multisyn, the HMM-based alignment transcription should be converted to the utterance structure. ASMO-449 is used to transliterate the Arabic text to English characters in the utterance files. Also, all items, features and relations are created to fit the same format of the utterance structure.

The speech wave files are coded using the Mel-frequency cepstrum coefficients (MFCC) and spectral coefficient and linear predictive coding (LPC). The voice building toolkit that comes with festival has been used for building our Arabic Emotional voice. The following steps were used:

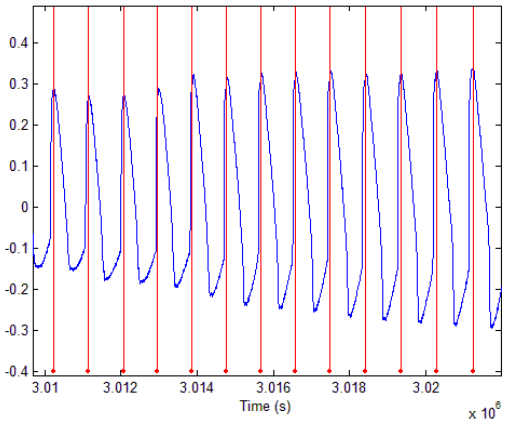
1. Generate power factors and wave file normalization
2. Using EGGs for accurate pitch marking extraction
3. Generating LPC and residuals from wave files
4. Generate MFCCs, pitch values and spectral coefficients

3.3 Pitchmarking

Pitch marking for unit selection voices is very important. Accurate estimation of pitch periods and pitch marks is necessary for pitch modification to assure optimal quality of synthetic speech. In our system we extracted pitchmarks from the EGG signal. . Matrix optimization processes have been used for pitch marking enhancement. The Edinburgh Speech Tools Library [18] was used for pitchmarking extraction. The function works by high and low pass filtering the signal using forward and backward filtering to remove phase shift. The negative going points in the smoothed differentiated signal, corresponding to peaks in the original are then chosen. The low pass filter and high pass filter cut-off frequencies have been chosen for the optimization process. Figure 2 shows an example of the default parameters of the pitchmarking application versus the optimized ones.



(a)



(b)

Figure 2. Close-up of pitchmarks in EGG signal. (a) Default pitchmarks extraction (b) optimized pitchmarks extraction

3.4 Challenges for building TTS Emotional voices

To have concatenative TTS voice with good quality usually you require large amount of data. To add emotion effects to the voice you would require sample recording for each emotion mode. Usually these emotion data sets are small in size compared with the available non-emotional data sets. Fortunately these recoded neutral data usually includes some phrases with blended emotions. Many segments that is uttered in a question and sad manner like some statements in stories and news that pronounced in a state of grief and sadness can be a precious soured for providing emotion data with low cost. The allocation of those emotional segments would require an emotionally speech recognizer. In this work we used the Emovoice tool.

4. EMOTIONAL SPEECH RECOGNITION

4.1 Emovoice Emotional Speech Recognizer

Emovoice is a framework for emotional speech corpus and classier creation and for offline as well as real-time online speech emotion recognition [14]. The system comes with a predefined two classification models; probabilistic naïve Bayesian and support vector machine (SVM) classifiers. Since an optimal feature set for speech emotion recognition is not yet established, Emovoice uses a composite feature sets in addition to some statistical and temporal features. The features used are logarithmised pitch, signal energy, Mel-frequency cepstral coefficients (MFCCs; 12 coefficients), the short-term frequency spectrum, and the harmonics-to-noise ratio (HNR). The resulting series of values are transformed to different views, and for each of the resulting series mean, maximum, minimum, range, variance, median, first quartile, third quartile and interquartile range are derived. These values constitute the actual features used [14]. Overall, The feature vector is containing 1302 features.

4.2 Proposed Emovoice Experimental setup

The NB classifier is very fast, even for high-dimensional feature vectors, and therefore especially suitable for real-time processing. However, it yields slightly lower classification rates than the SVM classifier which is a very common algorithm used in offline emotion recognition [19]. So, the SVM model is used with the complete feature set to build an offline emotional classifier.

The Emovoice model was trained on balanced data set consisting of one hour from each class (normal, sad and question). The normal uttered speech was manually selected from the 10 hours of the normal database. This trained model was evaluated using k-fold validation. With k equals to 10 the model achieved 95% accuracy for classifying the recordings emotion.

This model has been used to extract the emotional parts from the remaining normal uttered database. This process resulted in adding about half an hour classified from the sad and question emotions.

4.3 Festival Emotional Utterance

The utterance structures of both the utterances in the training databases and the target utterances –to be synthesized- had to be changed to carry emotional information. A new feature called emotion has been added to the word item type in the utterance structure. The speech database was also marked with that emotion feature. The emotion feature takes one of five emotional values. Table 1 shows the five emotional feature values and their meanings. These values have a great impact when calculating the target cost for each synthetic target units. The target cost function has been updated to take into consideration the emotional state between the target and candidate utterance.

Table 1. The five emotional feature values

Feature Value	**Meaning**
Normal	Normal state
Sad	Sad emotion state
Csad	Classified as sad from the normal database
Question	Question emotion sate
Cquestion	Classified as question from the normal database

5. PROPOSED EMOTIONAL TARGET COST

In unit-selection text-to-speech (TTS) systems, the task of unit selection is implemented by finding a sequence of database units that minimize the cost function. The cost function measures the distortion of the synthesized utterance, and is a summation of two sub-cost functions: a target cost, which describes the difference between the target segment and the candidate segment, and a concatenation cost, which reflects the smoothness of the concatenation between selected segments [14]. The target cost is a weighted sum of functions that check if features in the target utterance match those features in the candidate utterance and is computed for each unit in the candidate list. Each function returns a penalty if a feature does not match or if a penalty feature is set in the candidate. In other word, these differences are the p target sub-costs, $C_j^t(t_i, u_i)$ $(j = 1,2,3, \dots p)$ The target cost, given weights w_j^t for the sub-costs, is calculated as follows:

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i)$$

Instead of having a separate voice for each emotion, an emotional target cost module can be added to the target cost function to find the most suitable units for synthesized speech with the desired target emotion. The standard target cost in festival does not contain any computation differences for emotional state of the utterance. A new algorithm for computing the emotional target cost $(C_{emo}^t(t_i, u_i))$ is proposed. The algorithm is designed to work for blending database. For the desired emotional target utterances the emotional target cost $C_{emo}^t(t_i, u_i)$ is defined to be

$$C_{emo}^t(t_i, u_i) = \begin{cases} 0 & \text{Min Cost} \\ 0.5 & \text{Medium Cost} \\ 1 & \text{Max Cost} \end{cases}$$

The pseudocode of the proposed algorithm is described in figure 3. The algorithm in general favors units that are similar in the emotional sate or classified to be emotional in a three stages of penalties. The tuning weighting factor of the emotional target cost is optimized by try-and-error to find appropriate value.

```
Can_Emo  => Emotion of candidate unit
Tar_Emo  => Emotion of target unit
[Emo]*   => Classified as [Emo]

If (Tar_Emo == "sad" OR Tar_Emo ==
"Question") THEN
    If (Tar_Emo == Can_Emo) THEN
        Return NO_PENALTY
    Else
        If(Can_Emo == Tar_[Emo]*) THEN
            Return MID_PENALTY
        Else
            Return HIGH_PENALTY
Else
    If(Can_Emo =="normal" ) THEN
        Return NO_PENALTY
    Else
        Return HIGH_PENALTY
```

Figure 3. The proposed target cost algorithm

6. EVALUATION

Evaluation of emotional speech synthesizers is one of the biggest challenges in expressive speech technology research. The emotional categories are quite fuzzy in their definitions, and different researchers use different sets of emotions. The accuracy of the developed emotional synthesis was assessed in two major set of evaluation. The first one is listening tests. Using listening test, employing number of selected background listeners and listening environments is the most important issues in testing the performance and quality of output in an emotional synthesizer. The second one is using automatic emotion classification system (Emovoice). The underlying assumption is that an emotion synthesis system is of high quality, if the intended emotion can be predicted correctly by an emotion identification system that is trained on human voices [19].

6.1 Experimental setup

To evaluate the system, the emotional classifier in Emovoice is used across the produced utterance. Two types of listening tests were performed in order to evaluate the system perceptually. First, in order to evaluate the intelligibility of speech, a specific intelligibility test was performed. In this test, the listener was allowed to listen to the samples only once, and was then asked to type in what they heard. Word error rates (WER) of the answers were evaluated. Also, a subjective listening test is employed. The listeners are asked to rate the quality of the output voice. Two different ratings are used for naturalness and emotiveness. Participants were asked to give ratings between 1 and 4 for poor, acceptable, good and excellent respectively. The participant can listen to the sentences multiple times.

Six sentences were synthesized for each emotional state. The sentences were chosen to keep a certain average number of joins. The Sentences were from the news websites, usual conversations and the holy Qur'an. The total number of sentences is 18. The actual experiment took place on a computer where subjects had to listen to the synthesized sentences using headphones.

6.2 Results

The confusion matrix of the classifier output emotion and the target emotional state of the utterance is shown in Table 2. The overall classification accuracy exceeds 80%. For perceptual test, the experiment had 15 participants from which 6 were female and 9 were male. All of them are native Arabic speakers and postgraduate university students between the age of 22 and 33.

Table 2. Classification confusion matrix for test sentences

Classified As→	Normal	Sad	Question
Normal	5	1	0
Sad	0	6	0
Question	1	0	5

The participants were asked to write down what they heard. The Word Error Rate (WER) is then computed. The WER calculation is done on a normalized Arabic text by;

$$WER = \frac{E}{N}$$

Where E is number of wrong deleted, inserted or recognized words. N is the total number of words in a test sentence. The results of WER for different emotions are shown in Figure 4.

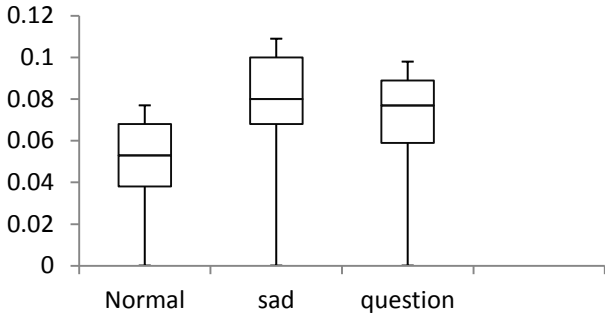


Figure 4. WER Mean and variance for different emotions

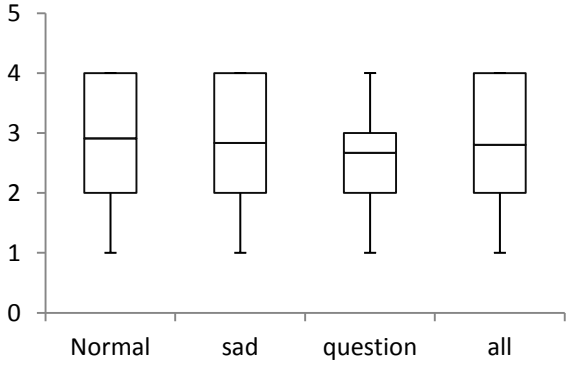
The listeners also rated the naturalness and emotiveness of test sentences. The descriptive statistics of the results are shown in Table 3 and Table 4. The results of mean and variance study for naturalness and emotiveness ratings also shown in Figure 5.

Table 3. Descriptive Statistics of the Naturalness

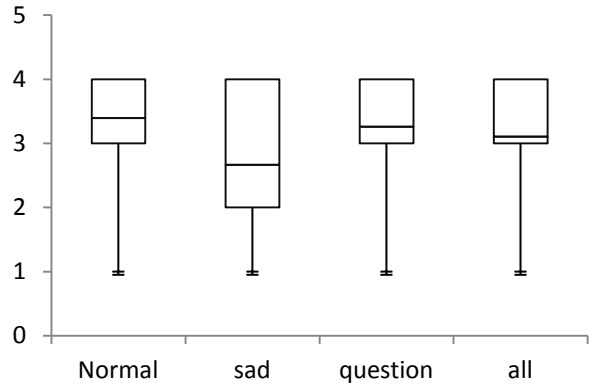
	Rating Mean	Standard Deviation
Normal	2.91	1.01
Sad	2.83	1.03
Question	2.67	0.98
All	2.8	1.01

Table 4. Descriptive Statistics of the Emotiveness

	Rating Mean	Standard Deviation
Normal	3.39	0.78
Sad	2.67	1.13
Question	3.26	0.86
All	3.11	0.98



(a) Naturalness



(b) Emotiveness

Figure 5. The means and variance of the ratings according each emotion

6.3 Discussion

The evaluation results using the emotional classification system show that the emotional characteristics of the synthesized speech are well recognized. But, this can't give a complete indication on the perceptual flavor of the output. In general, the hypothesis that more units from a given emotion database results in more accurate classification was supported.

The listening tests can give a more clear view on the results. First, the WER for different emotion styles are very small with maximum average of 8% in sad emotion. The participants were able to recognize the output sentences correctly to good extent. It is clear from the rating results that the naturalness and emotiveness are more than acceptable and reach good level for some emotions. From the descriptive statistics and variance analysis in Tables 1,2 and Figure 4. It is clear that the overall mean of naturalness ratings is 2.8 which approach a good quality naturalness. This does not deny the existence of low ratings outliers. The emotiveness ratings gave more promising results. It shows higher mean ratings value for all emotions in addition to lower variance for most cases. This gives us confidence on the results. The average ratings of overall emotiveness are 3.11 which indicate good emotive state of the synthesized speech. Also one of the interesting findings is the positive correlation between naturalness and emotiveness ratings. The naturalness and emotiveness ratings for question emotion sentences has lower mean value and high variance which means that they were not –to some extent- recognized as natural human speech. The rapid pitch variations of the questions emotion characteristics may stand behind this, but it stills acceptable according to the statistical measures.

7. CONCLUSIONS AND FUTURE WORK

The main goal of this research was to develop an Emotional Arabic TTS voice. This research focused on three important emotional sates; normal, sad and questions. Voice building is one of the vital processes in unit selection synthesis. The size of the used recorded speech database is the main critical factor for the quality of the produced voice. The using of emotion speech classification system (Emovoice) was very useful in increasing the emotional database size. This is in turn reflects positively on the quality of the output speech. Especially for blending approach it is recommended to increase the duration of acted or real emotional utterances in the RDI Arabic speech database. According to the different tests performed on the system, it shows promising results. At most the participants feel acceptable natural voice with clear good emotive state. However the work done for accurate pitch-marking, some further enhancements needed especially for question speech utterances.

8. ACKNOWLEDGEMENTS

Special thanks for the Research and Development Company (RDI) for providing the TTS database used in this research.

9. REFERENCES

[1] M Montero, J M Gutierrez-Arriola, S Palazuelos, E Enriquez, S Aguilera, J M Pardo. 1998. Emotional speech synthesis: from speech database to TTS. ICSLP

[2] Schröder, M. (2001). Emotional speech synthesis: A review. In Eurospeech 2001 Scandinavia. Proceedings of the 7th european conference on speech communication and technology, 2nd interspeech event. (pp. 561-4). Aalborg, Denmark, September 3-7, 2001.

[3] Roberto Barra-Chicote, Junichi Yamagishi, Simon King, Juan Manuel Montero, and Javier Macías Guarasa. 2010. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech, Speech Communication 52(5):394-404

[4] Marc Schröder. Expressive Speech Synthesis: Past, Present, and Possible Futures. 2009. Affective Information Processing, chapter 7.

[5] Marc Schröder. Emotional Speech Synthesis for Emotionally-Rich Virtual Worlds. 2003. the 8th International Conference on 3D Web Technology (Web3D)

[6] Brigitte Krenn, Hannes Pirker, Martine Grice, Paul Piwek, Kees van Deemter, Marc Schröder, Martin Klesen, and Erich Gstrein. Generation of multimodal dialogue for net environments. In Proceedings of Konvens, Saarbr¨ucken, Germany, 2002. URL http://www.ai.univie.ac.at/NECA.

[7] Irene Albrecht, J¨org Haber, Kolja Khler, Marc Schröder, and H.-P. Seidel. "May I talk to you? :-)" – Facial animation from text. In Proceedings of Pacific Graphics 2002, pages 77–86, 2002.

[8] Taylor,P. A., Black, A. W., & Caley, R. (1998) The architecture of the festival speech synthesis system. In The Third ESCA Workshop in Speech Synthesis, pages 147-151, Jenolan Caves, Australia

[9] Black, A.W., 2003. Unit selection and emotional speech. In: Proc. EUROSPEECH 2003, pp. 1649–1652.

[10] Alan W. Black. 2002. Perfect synthesis for all of the people all of the time. IEEE TTS Workshop 2002.

[11] Robert A. J. Clark, Korin Richmond, Simon King. 2007. Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication, 49(4):317-330.

[12] P. Taylor, A. Black, and R. Caley. 1998. The architecture of the festival speech synthesis system. In 3rd ESCA Workshop on Speech Synthesis, pages 147--141, Jenolan Caves, Australia.

[13] Wael Hamza and Mohsen Rashwan. 2000. "Concatenative Arabic speech synthesis using large

database", In Proceedings of ICSLP2000, vol. 2, pages 182-185, Beijing, China.

[14] Yong Zhao, Peng Liu, Yusheng Li, Yining Chen and Min Chu. 2006. Acoustics, Speech and Signal Processing. ICASSP 2006 Proceedings. 2006 IEEE International Conference on (Volume:1)

[15] Maria Assaf, Harald Berthelsen and Beata Megyesi. (2004). "A Prototype of an Arabic Diphone Speech Synthesizer in Festival". Msc Thesis.

[16] Al-Haj, H., Hsiao, R., Lane, I., Black, A., and Waibel, A. "Pronunciation Modeling for Dialectal Arabic Speech Recognition" ASRU 2009, Merano, Italy.

[17] Hassan Al-haj, Roger Hsiao, Ian Lane and Alan W. Black. (2009). Pronunciation Modeling for Dialectal Arabic Speech Recognition. ASRU, page 525-528.

[18] Anumanchipalli, G., Prahallad, K., Black, A. 2011. Festvox: Tools for Creation and Analysis of Large Speech Corpora. in Proceedings of Very Large Scale Phonetics Research, UPenn, 2011.`

[19] Stefan Steidl, Tim Polzehl, H. Timothy Bunnell, Ying Dou, Prasanna Kumar Muthukumar, Daniel Perry, Kishore Prahallad, Callie Vaughn, Alan W. Black, and Florian Metze, Emotion Identification for Evaluation of Synthesized Emotional Speech Speech Prosody 2012, Shanghai, China.