OPTIMIZATION OF COST FUNCTION WEIGHTS FOR UNIT SELECTION SPEECH SYNTHESIS USING SPEECH RECOGNITION

Miran Pobar, Sanda Martinčić-Ipšić*, Ivo Ipšić**

Abstract: A well known problem in unit selection speech synthesis is designing the join and target function sub-costs and optimizing their corresponding weights so that they reflect the human listeners' preferences. To achieve this we propose a procedure where an objective criterion for optimal speech unit selection is used. The objective criterion for tuning the cost function weights is based on automatic speech recognition results. In order to demonstrate the effectiveness of the proposed method listening tests with 31 naive listeners were performed. The experimental results have shown that the proposed method improves speech quality and intelligibility. In order to evaluate the quality of synthesized speech the unit selection speech synthesis system is compared with two other Croatian speech synthesis systems with voices built using the same recorded speech corpus. One of these voices was built with the Festival speech synthesis system using the statistical parametric method and the other is a diphone concatenation based text-to-speech system. The comparison is based on subjective tests using MOS (mean opinion score) evaluation. The system using the proposed method used for cost function weights optimization performs better than other compared systems according to the subjective tests.

Key words: *Speech synthesis, statistical parametrical synthesis, unit selection, weight tuning*

Received: October 6, 2011
Revised and accepted: September 17, 2012

1. Introduction

Currently two different corpus based methods dominate in speech synthesis research. The unit selection method is based on selecting and concatenating units of natural speech from the available corpus. If there is more than one instance of each unit spoken in different styles, the system can choose the sequence of units that best conforms to desired prosody and has the least audible joins. In most

*Miran Pobar, Sanda Martinčić-Ipšić, Ivo Ipšić
University of Rijeka, Department of Informatics, Radmile Matejčić 2, 51000 Rijeka, Croatia, E-mail: {mpobar,smarti,ivoi}@uniri.hr

systems the selection is guided by two cost functions, join and target cost, and the string of acoustic units with minimal total cost is selected. The cost functions may include contributions of several features calculated as sub-costs. Ideally, these functions should predict the human perception of quality of the resulting speech. To this end, an optimal set of join and target sub-costs and their corresponding weights should be determined. Various approaches based on perceptual evaluation involving human listeners have been proposed e.g., [7, 10, 11, 18, 23]. The evaluation process can be both time consuming and expensive to conduct, which is especially a problem if it needs to be done many times during the course of system development, e.g. when a voice using a new speaker is developed. Approaches based on objective measures have also been proposed [2, 9, 13], with an advantage that the optimization can be repeated easily as many times as necessary, with consistent and comparable results between runs. However, an objective measure may not be a good predictor of human perception. Objective measures are sometimes designed for tuning a specific part of the cost function, and cannot be generalized for tuning all the weights, complicating the process. For example, in [9] the target cost weights are set according to the criterion of minimal cepstral distance between the natural reference target utterances and the synthesized utterances. The same measure would be difficult to apply to join cost, and may also penalize utterances that a human would rate as good, but are acoustically further from the reference waveform.

Quality of speech produced by unit selection can vary widely, depending on availabilty of units in the database and the selection algorithm itself. Variants of unit selection synthesis have been implemented in systems such as ν-talk [19], CHATR [9], and Festival [8].

Statistical modeling of speech has recently been successfully applied in speech synthesis, leading to the statistical parametric speech synthesis. The method is based on parametrization of speech that can be both reversed and modeled. A set of models is trained on examples of natural speech. At synthesis time, these models can produce speech parameter vectors from which speech waveform is generated. Typically, hidden Markov model (HMM) formalism is used along with mel-frequency cepstral coefficients (MFCC) [24, 4], but other parameters such as formant trajectories have been used as well [1]. The reported quality of such systems is generally very good [3], although it still has some drawbacks compared to unit selection systems, especially the buzziness of generated speech resulting from the filtering process used to generate the waveforms.

In this work, the possibility of using automatic speech recognition (ASR) results as the optimization criterion for weight tuning is explored. The criterion was tested on a Croatian speech synthesis system, and the optimization results are presented. A formal evaluation listening test was conducted to validate the approach. A statistical parametrical system was also built using the same speech corpus that was used for the unit selection voice so the systems can be compared. The two systems were also compared with a diphone concatenation system based on the PSOLA [15] method, also built using the same corpus. A subjective evaluation of the systems was conducted using the mean opinion score (MOS) scale.

In the next section, the basic unit selection speech synthesis method is presented. Croatian speech synthesis and recognition systems are described in Sec-

tion 3. In Section 4 the proposed criterion for weight tuning and the tuning procedure are presented. Section 5 presents the results of weight tuning and the perceptual evaluation. In Section 6, comparison of the 3 systems and the results of subjective system evaluation are presented. Finally some conclusions and suggestions for future work are given.

2. Unit Selection Speech Synthesis

The unit selection speech synthesis method is based on concatenation of recorded segments, or units, of natural speech stored in a corpus. A general unit selection scheme was proposed in [9]. The input to the synthesizer is a target specification $S = \{s_1, s_2, \ldots, s_N\}$, which is a sequence of N units described with feature vectors. The goal of unit selection is to find the optimal sequence of units $U = \{u_1, u_2, \ldots, u_N\}$ in the corpus corresponding to the specification S. The choice of best unit sequence is presented as the problem of finding the path with a minimal total cost through a network where each node is a unit in the database, and the edges are possible joins. Two cost functions are defined: target and join cost. The target cost $T(u_i, s_t)$ is a difference measure between a unit u_t in the corpus and a target unit s_t, i.e. the desired rendition of this unit. The target cost is effectively the cost of a given node. The join cost $J(u_t, u_{t+1})$ is a measure of how perceptible is the join between two consecutive units u_t and u_{t+1}, and corresponds to the cost of the edge from unit u_t to u_{t+1}. Both target cost and join costs are defined as weighted sums of p and q sub-costs, where p and q depend on the number of features that contribute to each cost:

$$T(u_t, s_t) = \sum_{j}^{p} w_j^t T_j(u_t, s_t) \tag{1}$$

$$J(u_t, u_{t+1}) = \sum_{j}^{q} w_j^j J_j(u_t, u_{t+1}) \tag{2}$$

and the factors w^t and w^j are target and join weights. The target and join sub-cost functions measure contributions of individual features to the cost, usually in a form of difference of feature values. These features may be a combination of linguistic and acoustic features for the join cost, and linguistic features for the target cost because acoustic features usually are not available for the target specification. The weights determine the relative importance of sub-costs and are an important part of cost functions design, as they should reflect the subjective perception of human listeners. The total cost of a sequence of units is a sum of target and join costs for N units in the utterance:

$$C(U, S) = \sum_{t=1}^{N} T(u_t, s_t) + \sum_{t=1}^{N-1} J(u_t, u_{t+1}) \tag{3}$$

and the optimal sequence is

$$\hat{U} = \arg\min_{U} C(U, S). \tag{4}$$

The Viterbi dynamic programming algorithm can be used to find the optimal path through the network. At synthesis time, the system searches the corpus to find a sequence of units that matches the desired phonetic string and concatenates the corresponding waveforms.

3. Description of Systems

3.1 Corpus and phoneset

The same speech data from the VEPRAD [12] radio news and weather report corpus was used to build the voices. The VEPRAD corpus is a multi-speaker database, containig speech from 11 male and 14 female professional speakers. From this corpus a subset from a single male speaker with the most data was selected for building the voices. This subset consists of about 2 and a half hours of transcribed speech with word level textual transcriptions. The size of this set is 267 MB. Phone level segment labels were obtained automatically using HMM speech recognition in forced alignment mode. The recognition is done using the HTK toolkit, with monophone HMMs trained on the same speech corpus that is used in synthesis. The ASR system was trained on the whole corpus speech with over 208K uttered words and 15K unique words [12].

The phoneset that was used to build the voices consists of 30 standard phonemes of the Croatian language, 5 accented forms of vowels, the syllable-forming /r/, and the silence phoneme (37 phonemes in total). In the Croatian language, grapheme-to-phoneme conversion is mostly straightforward, with one to one mapping in a majority of cases. A basic set of manually produced mapping rules for 30 graphemes is enough to cover unknown words. Better quality of speech can be obtained with a more extensive set of rules that take into account various sound changes occurring in speech, so they were adapted for these voices as well [16]. A lexicon consisting of about 10000 entries with phonetic transcriptions of words with stress position information is used in conjunction with the rules, as they do not predict stress placement.

3.2 Statistical parametric voice for festival

The statistical parametric voice was built using the Festival and Festvox [5] tools, with the clustergen [4] statistical parametric synthesis module.

Each context dependent phoneme is modeled using a three state HMM with the set of parameters consisting of 24 MFCCs and log F0, extracted at 5 ms intervals. Duration of each HMM state is predicted using a separate classification and regression tree (CART) trained on the data used to build the voice, utilizing linguistic features extracted from input text as questions at tree nodes.

To build the Croatian voice the phoneset was adapted to the Festival system. For each phoneme linguistic features that are used in CART questions were defined according to the Croatian phonetic system [20], such as whether the phone is a vowel or a consonant, its place of articulation, length, voicing etc. Grapheme to phoneme rules and lexicon were converted from Matlab and Perl to the Scheme language for Festival.

To train the context-dependent HMMs acoustic features (MFCCs, voicing information and log F0) need to be extracted from the data first. This is done using tools provided with Festvox, as well as preparation of utterance files with linguistic features. The next phase involves training the actual HMM models. The HMM training stage requires HMM state level segmentation of speech and as the corpus was segmented to the phone level the data needed to be segmented again. The state segment labels were generated using the Festvox scripts and the EHMM recognizer provided with the Festival tools. Finally, the HMM parameters and duration CART trees were estimated from the aligned utterances and extracted features. The resulting voice takes up 10.2 MB.

3.3 Diphone voice

A diphone concatenative voice was developed for a custom TTS system. Only one instance of each diphone was kept, resulting in 923 diphones out of theoretically possible 37*37=1369 for the chosen phoneme set. Diphones from word middles were preferred and the instance with duration closest to the average was selected. Along with the waveform, the glottal closure instances computed using the DYPSA algorithm [17] were kept. The concatenation procedure was implemented in Matlab and was based on the PSOLA algorithm [15]. No pitch modification is done at synthesis time so the resulting prosody depends only on the available diphones.

3.4 Unit selection voice

The unit selection voice built for a custom system follows the general unit selection framework described in [9]. In this system, the diphone was chosen as the fundamental acoustic unit. First, the unit database was populated with a number of units for each diphone class. With each unit a number of features were stored (left and right phone identity, beginning and ending times of both phones, $F0$ contour, glottal closure instants, 12 MFCCs, log energy, first and second formant frequency at concatenation point and a context identifier).

$F0$ contours were extracted in 10 ms frames using the RAPT [22] algorithm and smoothed using a three-point running median filter. For unvoiced regions, the missing $F0$ values were inserted using linear interpolation from neighboring values. The glottal closure instants were detected using the DYPSA [17] algorithm and the formant contours were extracted using the Snack Sound Toolkit [21]. Twelve MFCC coefficients and log energy were extracted in 16 ms frames with 8 ms overlap. A unique context identifier is also stored with each unit so the units that were originally joined in natural speech may be identified when calculating the join cost.

For the target function, the normalized Euclidean distance between the durations of phonemes and corresponding phoneme class mean durations was used, as shown in (5):

$$T(u) = \sqrt{\frac{\left(d_l - \overline{d_l}\right)^2}{\sigma_l^2} + \frac{\left(d_r - \overline{d_r}\right)^2}{\sigma_r^2}}\,, \tag{5}$$

where u is the current unit (diphone), d_l and d_r the durations of left and right phoneme in the unit u respecively, $\overline{d_l}$ and $\overline{d_r}$ mean durations of left and right phoneme classes and σ_l and σ_r are the standard deviations of left and right phoneme class durations.

Join cost is a weighted sum of absolute difference of $F0$ value at the point of concatenation, absolute differences of first and second formant frequencies and the distance of MFCC vectors. MFCC distance is computed as the Euclidean distance between MFCC vectors of the first frame of next diphone and one frame after the last in the current diphone. If two diphones that are adjacent in the original recording are considered, the MFCC distance becomes zero. When diphones come from different utterances, the one frame overlap ensures that lower cost is assigned to the unit with spectral characteristics similar to the continuation of the current diphone, and not the current diphone itself. This is particularly to account for phones with abrupt spectral changes, e.g. plosives.

Join cost $J(u_t, u_{t+1})$ between units u_t and u_{t+1} is defined as:

$$J(u_t, u_{t+1}) = w_f \left(|F0_t - F0_{t+1}|\right) + w_m \left(|F1_t - F1_{t+1}| + |F2_t + F2_{t+1}|\right) + w_c \sqrt{\sum_{i=1}^{12} \left(c_{it} - c_{it+1}\right)^2}, \quad (6)$$

where w_f, w_m and w_c are the $F0$, formant and MFCC cost weights, $F0$ is the fundamental frequency, $F1$ and $F2$ first and second formant frequencies and c is the MFCC vector.

The total cost $S(u_{t+1})$ of choosing the unit u_{t+1} is

$$S(u_{t+1}) = w_t T(u_{t+1}) + w_j J(u_t, u_{t+1}), \quad (7)$$

where w_t and w_j are target and join cost weights.

The weights were trained according to the procedure described in the next section. The unit sequence with the minimal total cost over the whole utterance is found using the Viterbi algorithm and the speech waveform is generated using the overlap-and-add technique. No extra signal processing was used. The system was implemented in Matlab, using Voicebox [6] for MFCC extraction and the RAPT algorihm for $F0$ estimation.

3.5 Speech recognition system

The Croatian ASR system is based on continuous hidden Markov models (HMM) of monophones and triphones trained with the HTK Toolkit [25]. The monophone models with continuous Gaussian output probability functions were trained for the 30 standard Croatian phonemes and 4 additional models for silence, breathing (inspiration) sound, mispronounced words, hesitations and noise. The initial training of the Baum-Welch algorithm on HMM monophone models resulted in a monophone recognizer, which was used for the automatic segmentation of the speech signals. The automatically segmented speech database was used to model triphone HMMs. The resulting phone level segments for the selected speaker were used for

the speech synthesis system as well. The further training of the ASR system resulted in triphone acoustic models with continuous density output functions (one to twenty mixture Gaussian density functions), described with diagonal covariance matrices. The state tying was performed due to the lack of the acoustic material using proposed Croatian phonetic rules [14]. For speech recognition the speech signal feature vectors consist of 12 mel-cepstrum coefficients and their derivatives and acceleration. The feature coefficients were computed every 10 ms for a speech signal frame length of 20 ms. In the ASR system, a backoff bigram language model is used with estimated perplexity of 16.94. The achieved word error rate (WER) for the weather domain recognition task was below 5% [14]. For the purpose of this work, a unigram language model with 14620 words and uniform word probabilities was used.

4. Weight Tuning

The proposed objective criterion for optimizing the cost functions weights is the word correctness metric commonly used in evaluation of ASR system performance. Word correctness WC of a sentence is defined as

$$WC = 100\%(1 - \frac{W_S + W_D}{N}), \tag{8}$$

where W_S, and W_D are substituted and deleted words, and N is the total number of words. W_S, and W_D are computed using the Levenshtein distance between the reference and recognized sentences. The goal of optimization is thus to find the weights that maximize the word correctness score for the sentences in the training set. To tune the weights a limited search of the weight space was done using a procedure similar to the approach described in [9] employing the proposed objective criterion. However, for this task a number of other optimization techniques could be used instead. In the experiment, the set of weights to be optimized consists of join and target cost weights and the three join cost function sub-weights (eqs. 6 and 7). The same set of weights was used for all units. First, a set of 100 words was chosen randomly from the dictionary of the speech recognition system. The chosen words were not present in the data used to build the unit inventory for speech synthesis. For each weight set, the words were synthesized in isolation and fed to the speech recognizer. In this case, the calculation of word correctnes becomes simply the percentage of correctly recognized words. The word correctness score was calculated and the weight set with the highest score was chosen as best. Using larger utterances instead of isolated words was first considered for the experiment but later rejected for two reasons. The first was to minimize the influence of the speech recognition system's language model and favor the acoustic model. The language model was reduced to a unigram model and unigram probabilities for all words were reset to a uniform value. This could also be done for larger utterances, so this reason was not very important. The other reason was to make evaluation easier, as listeners in a preliminary test reported it was difficult to concentrate and compare longer segments, especially if the quality within the utterance varied. In total, 48 weight sets were tried and produced word correctness scores ranging from 6 to 18%. To verify that the results were dependent on the chosen weights and are

not a consequence of the chosen words a new set of 100 randomly picked words was generated and the process was repeated for the same weight sets. The same weight sets generated the lowest and highest word correctness scores on both sets of words.

5. Perceptual Evaluation of Agreement Between Human Perception and ASR Results

A blind A-B preference listening test was conducted to verify the agreement of results obtained using the recognition results with human perception. The weight sets which produced the lowest and highest recognition results were used in the test (low and high set). From the 100 words used to select the best weight set, 17 were kept for the listening test. Each of these words was synthesized using the high and low weight sets, giving two variants for each of the selected 17 words. From the starting hundred words, five were correctly recognized for both weight sets and 12 were correctly recognized only using the weight set which gives the highest word correctness. All words that were correctly recognized using the low weight set were also correctly recognized when using the high weight set. The words that were not recognized correctly in neither one of two sets were not included in the test. It was expected that the listeners would prefer the version of the word that was correctly recognized by the ASR system. The evaluation was conducted using an interactive GUI application. For each word, labels A and B were randomly assigned by the application to waveforms from low and high weight sets. The listeners could play back the waveforms A and B as many times as needed in arbitrary order, and could choose options "prefer A", "prefer B" or "no preference". The tests were conducted in a relatively quiet office and all listeners used the same headphones. The evaluation normally took about 3 minutes per listener to complete. There were in total 31 naive listeners with no previous experience with speech synthesis, all native Croatian speakers, mostly graduate students of computer science.

The results of the preference test for each of the 17 words are shown in Fig. 1. The results show that the listeners had no preference between words generated using low and high weight sets for the first five words, while for the other twelve words the listeners preferred the words synthesized using the high weight set. First five words were correctly recognized using ASR, for both weight sets and thus it can be said that from the ASR perspective they cannot be distinguished. The human listeners seem to agree as they equally likely chose preference for words obtained with either of the weight sets. For words 6-17 the variants synthesized with the high weight set were correctly recognized using ASR, while the variants obtained using the low set were not. It can be said that the ASR “prefers” the variants of words obtained using the high set. In this case, the human listeners also agree, as they in higher proportion preferred the variant obtained using the high set. This result shows agreement of human perception and ASR results and justifies the use of ASR results as an approximation of human perception for weight tuning. Fig. 2 shows the cumulative preference percentages for words that were correctly recognized using both low and high weight sets, and for words correctly recognized only using the high weight set. As noted, for the first group where there was no

difference in ASR performance, the votes were divided between both variants, with 39% votes for waveforms generated using the high set, 22% with no preference and 39% for the low set. For the second group the majority of listeners preferred the waveforms generated using the high weight set (61% votes for high vs. 20% with no preference and 18% for the low weight set). The listening tests proved that the criterion function for weight set selection satisfies the goal, which is to improve the synthesized speech quality and intelligibility.

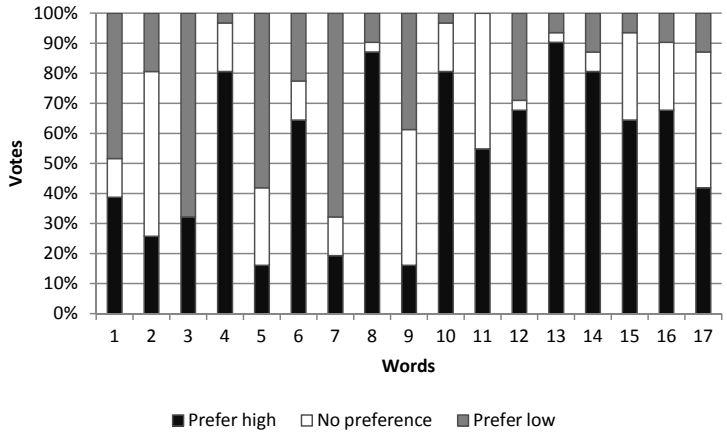

Fig. 1 *Listener preference percentages of base and optimized versions of synthesized words.*

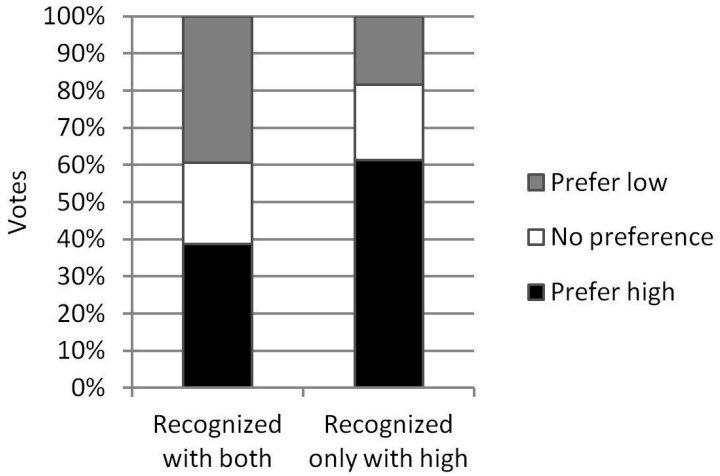

Fig. 2 *Cumulative listener preference percentages for groups of words. The left bar represents words that were correctly recognized with both low and high weight sets. The right bar represents words correctly recognized only with high weight set.*

6. Evaluation and Comparison

Preliminary evaluation using MOS (mean opinion score) was conducted for all developed voices. For each voice, two samples of synthesized speech were prepared. Text A was a synthesized text from the weather domain, using the same vocabulary as in the training corpus. The wording of the text was different from any utterance in the training corpus. Text B was from the news domain with a larger proportion of words not present in the lexicon. Both texts have three sentences, with total length of 34 and 37 words, respectively.

On a scale from 1-5, with 1 being the worst and 5 the best, the listeners evaluated the overall quality, intelligibility, naturalness and occurences of irregularities in the synthesized speech. The listeners also responded whether they thought the synthesized speech was acceptable for use in an automated information service over the telephone, with answers yes, yes with improvements to the system and no, corresponding to scores 5, 3 and 1. The samples were presented to the users using a web page and the results were collected using an electronic form. The participants listened to the samples in arbitrary order and could repeat the samples any number of times. Twelve listeners of both genders (mostly university students and staff) participated in the evaluation. Of those, 6 had previous contact with speech synthesis systems as participants in an unrelated evalutaion. The MOS evaluation results are presented as a box plot in Fig. 3. The medians are represented by solid bars across boxes that show the quartiles and the whiskers extend to 1.5 times the

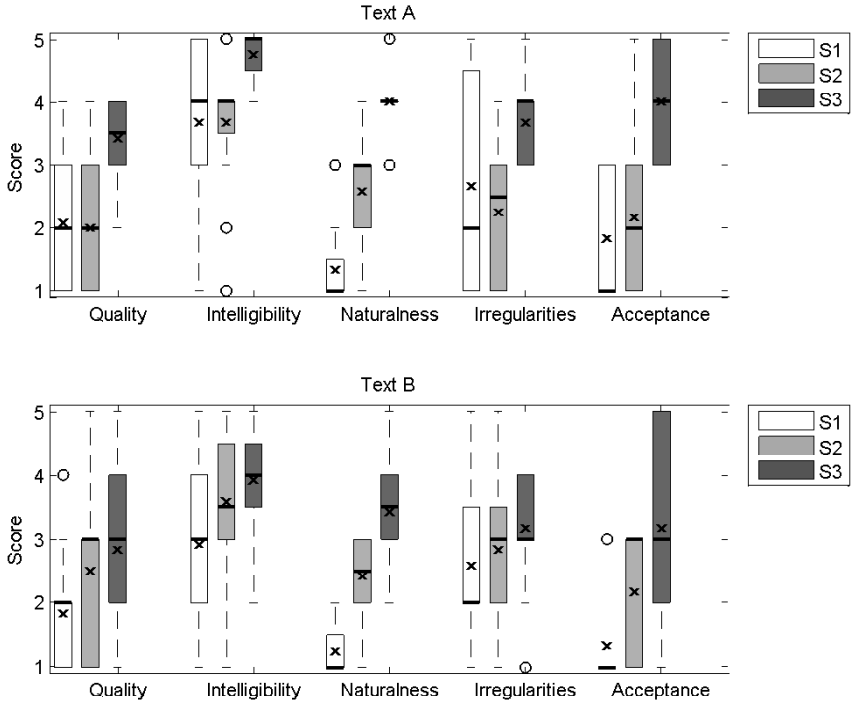

Fig. 3 *MOS results for text A (top) and B (bottom).*

inter-quartile range. Outliers are represented by circles and means by crosses.

In the following discussion, the voices described in sections 3.2, 3.3 and 3.4 are labeled S1, S2, and S3. For both texts, the unit selection system got the best scores for all questions. The join cost itself may choose the units that join well, but it only considers neighboring units so the overall prosody of the whole utterance may still be poor, which was reported by some listeners as a "singing" quality. A very simplistic target function used in the system S3 should thus be improved.

As expected, the scores of the system S3 were better for text A than for text B where more unknown words were present. The system S2 (the diphone system) performs about the same for both texts as it cannot take advantage of the fact that text A contained words present in the training corpus and the number of concatenations is always the same. The voice S1 also performed better on text A, but the difference was smaller. This suggests that the voice may generalize better outside the domain of the training corpus than the S3 voice, however the overall score was still lower.

Low scores for system S1, which is even outperformed by the diphone system in most cases, were somewhat surprising, considering other comparisons [3]. Particularly low score was given for its naturalness, which was expected due to the vocoding buzziness. Some participants noted that although the inteligibillity of the voice was excellent, the metallic character of the voice made it unsuitable for use in automated information services. 75% of users described system S1 as acceptable for this use for text A.

7. Conclusion and Future Work

The perceptual evaluation results support the use of speech recognition results as an objective criterion for optimizing the weights in the unit selection concatenative speech synthesis system. As the chosen optimization method searched over a rather small portion of the weight space, the obtained set was probably not optimal so other optimization techniques will also be tested in conjunction with the proposed criterion. Also, the same criterion could be used to train the weights individually for each phoneme class or phoneme group, as in multiple regression methods. Although in this work the proposed criterion was used to optimize a set of weights comprising mostly join cost function sub-costs, it could also be applied to the target cost. This is an advantage over some other objective measures that are tailored either to optimize the join cost or target cost functions, e.g. mel-cepstral distance between natural target utterance and synthesized utterance, as in [9]. A comparison of results obtained for target function using both approaches would be insightful.

Three voices using diphone concatenation, unit selection and statistical parametrical approach were built for the Croatian language using the same speech corpus. The systems were evaluated and rated on a MOS scale for quality, intelligibility, naturalness, frequency of speech irregularities and acceptance.

The unit selection system performed best with the mean quality scores of 3.42 and acceptance scores of 4 for text in domain of the training corpus. For text with unknown vocabulary, the unit selection system still performed best but with lower scores, as was expected.

To improve the quality of speech from the unit selection systems several modifications should be made. Unsupervised clustering of the diphone database using neural networks could produce a useful feature for total cost calculation. The features of diphones that are used at runtime to calculate the join or target costs could be used as inputs to the network, and the resulting clusters can be stored in the database with the diphones. At syntheisis time, the cluster to which the diphone belongs can be either used as an additional feature of the join or target function, or used to reduce the search space and thus time needed to find the optimal sequence by considering only diphones of the set clusters. The target function should be changed to accomodate the sentence level prosodic information. To this effect, the linguistic context information may be combined with statistical models of prosody trained from data. Also, for the optimization of the cost functions weights, a method which searches over a larger space of weights will be evaluated. Weight optimization using ASR with samples of phrases or sentences instead of isolated words will be explored, which will require careful preparation of listening tests. Using phrases or sentences longer term prosodic effects can be factored in the perceptual evaluation. However, a concern over using longer phrases is that any isolated error in a word can possibly influence the listeners' judgement too much, and that error may or may not be directly caused by the chosen set of weights. Using larger number of phrases could prevent the influence of such random errors influencing the score, however there should not be too many, as the listeners may become too tired and lose concentration.

Acknowledgments

Support for this work was provided by the Ministry of Science, Education and Sports of the Republic of Croatia (project number 318-0361935-0852).

References

[1] Acero A.: Formant Analysis and Synthesis Using Hidden Markov Models. In: EUROSPEECH'99, 1999, ISCA, pp. 1047–1050.

[2] Alıas F., Llora X.: Evolutionary Weight Tuning Based on Diphone Pairs for Unit Selection Speech Synthesis. In: EUROSPEECH-2003, 2003, pp. 1333–1336.

[3] Barra-Chicote R., Yamagishi J., King S., Montero J., Macias-Guarasa J.: Analysis of Statistical Parametric and Unit Selection Speech Synthesis Systems Applied to Emotional Speech. Speech Communication, **52**, 5, 2010, pp. 394–404.

[4] Black A.: CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling. In: INTERSPEECH-2006, 2006.

[5] Black A., Lenzo K.: Building Synthetic Voices. Language Technologies Institute, Carnegie Mellon University and Cepstral LLC, 2003.

[6] Brookes M., et al.: Voicebox: Speech Processing Toolbox for Matlab. World Wide Web, http://www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html 2000.

[7] Chu M., Peng H.: An Objective Measure for Estimating MOS of Synthesized Speech. In: Seventh European Conference on Speech Communication and Technology, 2001, pp. 2087–2090.

[8] Clark R., Richmond K., King S.: Festival 2-Build Your Own General Purpose Unit Selection Speech Synthesiser. In: Fifth ISCA Workshop on Speech Synthesis, 2004.

[9] Hunt A., Black A.: Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: ICASSP-96, **1**, 1996, pp. 373–376.

[10] Lee M.: Perceptual Cost Functions for Unit Searching in Large Corpus-Based Concatenative Text-to-Speech. In: Eurospeech, 2001, pp. 2227–2230.

[11] Lee M., Lopresti D., Olive J.: A Text-to-Speech Platform for Variable Length Optimal Unit Searching using Perception Based Cost Functions. International Journal of Speech Technology, **6**, 4, 2003, pp. 347–356.

[12] Martinčić-Ipšić S., Ribarić S., Ipšić I.: Acoustic Modelling for Croatian Speech Recognition and Synthesis. Informatica, **19**, 2, 2008, pp. 227–254.

[13] Meron Y., Hirose K.: Efficient Weight Training for Selection Based Synthesis. In: EUROSPEECH'99, 1999, ISCA, pp. 2319–2322.

[14] Meštrović A., Bernić L., Pobar M., Martinčić-Ipšić S., Ipšić I.: A Croatian Weather Domain Spoken Dialog System Prototype. Journal of Computing and Information Technology, **18**, 4, 2011, pp. 309–316.

[15] Moulines E., Charpentier F.: Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication, **9**, 5-6, 1990, pp. 453–467.

[16] Načinović L., Pobar M., Ipšić I., Martinčić-Ipšić S.: Grapheme-to-Phoneme Conversion for Croatian Speech Synthesis. In: MIPRO 2009, 2009, **3**, pp. 318–323.

[17] Naylor P., Kounoudes A., Gudnason J., Brookes M.: Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm. IEEE Transactions on Audio, Speech and Language Processing, **15**, 1, 2007, pp. 34–43.

[18] Peng H., Zhao Y., Chu M.: Perpetually Optimizing the Cost Function for Unit Selection in a TTS System With One Single Run of MOS Evaluation. In: Seventh International Conference on Spoken Language Processing, 2002, pp. 2613–2616.

[19] Sagisaka Y., Kaiki N., Iwahashi N., Mimura K.: ATR ν-Talk Speech Synthesis System. In: ICSLP'92, 1992, ISCA, pp. 483–486.

[20] Silić J., Rosandić D.: Osnove fonetike i fonologije hrvatskog književnog jezika. Školska Knjiga Zagreb, 1983.

[21] Sjolander K., Beskow J.: Wavesurfer-an Open Source Speech Tool. In: ICSLP-2000, **4**, 2000, pp. 464–467.

[22] Talkin D.: A Robust Algorithm for Pitch Tracking (RAPT). Amsterdam, NL: Elsevier Science, 1995.

[23] Toda T., Kawai H., Tsuzaki M.: Optimizing Integrated Cost Function for Segment Selection in Concatenative Speech Synthesis Based on Perceptual Evaluations. In: EUROSPEECH-2003, 2003, pp. 297–300.

[24] Yamagishi J., Zen H., Toda T., Tokuda K.: Speaker Independent HMM-Based Speech Synthesis System-HTS-2007 System for the Blizzard Challenge 2007. In: BLZ3-2007, 2007.

[25] Young S., Evermann G., Gales M., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.: The HTK Book Version 3.4. Cambridge University Engineering Department, 2006.