

Definition of a Training Set for Unit Selection-Based Speech Synthesis

Karlheinz Stöber⁺, Petra Wagner⁺, Esther Klabbers^{*}, and Wolfgang Hess⁺

⁺ Institut für Kommunikationsforschung und Phonetik
Universität Bonn, Germany

^{*} IPO, Center for User-System Interaction
{kst, pwa, wgh}@ikp.uni-bonn.de ;
E.A.M.Klabbers@tue.nl

Abstract

The definition of cost terms in unit selection based synthesis is a difficult task. Usually cost terms are based on common phonetic knowledge of the developers and subsequent perceptual experiments. The dataset used for supervised learning, well known from pattern recognition, could be a useful way to arrive at a more formal analysis of the different factors influencing the selection of units.

As a first step toward this aim we present an objective distance measure which is used to sort the units contained in the corpus in relation to a given natural unit and prove its relevance to human perception. To avoid too much attention of the listeners to discontinuities caused by concatenation, we will also present a waveform-based smoothing algorithm.

It is experimentally shown that the sorting criterion and the human perception match in most cases. Furthermore it can be detected that similarity between natural and synthetic speech is better if phoneme based units are used, but naturalness increases with the concatenation of larger units.

1. Introduction

Recently non-uniform unit selection based speech synthesis has become popular, because the resulting speech often sounds more natural than that of traditional synthesis approaches. Most of the unit selection algorithms are based on a cost function as defined in [1]. Such a cost function consists of a set of terms which can be subdivided into two groups, namely *continuity distortion* and *unit distortion*. The task of the unit distortion terms lies in comparing the desired properties of the synthetic utterance with the (natural) unit properties contained in the synthesis corpus. This comparison is based on a numeric value calculated from non-numeric input parameters. Each term of the cost function is weighted by a scalar which adjusts the importance of the regarded feature as well the value range.

Both the search for adequate cost terms and proper weights are tasks often approached manually. There are some methods that optimize weights using an automatic training procedure, e.g. [1]. A serious problem of such methods is their computing complexity often growing exponentially with the number of cost terms.

A training dataset as used by supervised learning could first provides hints, as to which properties are useful in cost terms. It could then provide the possibility of proving the quality of the unit selection relative to a test

dataset. Finally it allows the automatic training of different prediction strategies such as HMM and ATNN. Such a training data set consists of a set of training patterns. Each training pattern is built from input parameters and assigned output parameters. Currently we regard the phonetic transcription of a natural utterance together with the regarded unit as the input parameters. The output parameter is the index of a unit in a synthesis corpus. This unit should be the perceptually most similar one compared with the regarded unit in the natural utterance.

This work concentrates on finding an optimal output parameter set. An objective distance measure between speech units is the prerequisite for solving this task, since a solution based on listening experiments would be impossible. Based on our results, the definition of criteria for defining input parameters which are better suited than a phonetic transcription will be a future goal. The ultimate step will then be the application of the training pattern to classification mechanisms. However, finding an objective distance measure and validating its quality is not a trivial task.

We assume that the speech corpus used for the synthesis contains all segmentally important units but is still finite. It follows that there must be a sequence of units producing the synthetic utterance considered best among all other possible sequences by a large group of human listeners. Starting from that point of view, all that is needed is a large set of perceptual experiments providing us with the desired training dataset. Since their complexity is exponential such experiments can unfortunately neither be realized with humans nor with computers. Another problem is the definition of naturalness. Especially when it comes to prosodic variants, different humans tend to have different preferences. This even holds for natural speech, as shown in [2]. Until today, it seems impossible to describe and compute all dimensions of naturalness necessary to define an objective measure.

Since all these problems appear unlikely to be solved in the near future, we concentrate on the optimization of unit distortion terms. Thereby ignoring continuity distortion, we can simplify the problem computationally, but are still in need for a strategy to avoid discontinuities at concatenation boundaries. Quality measurements for speech transmission are a well-established research area. In this field, good results have been achieved for psycho-acoustic measures [3]. Therefore, it can be assumed for these measures to be applicable to synthesis research as well. We extend the PEMO measure used in [3] toward handling units of different length and utilize it for a spectrally based minimization of the unit distortion. In

order to reduce the discontinuities at concatenation boundaries we develop a novel strategy for spectral smoothing.

The results show that listeners often agree with the preferences computed by the objective distance measure, but a high degree of objectively shown similarity does not imply a high perceived naturalness.

2. Application of the objective distance measure

Cost-based unit selection regards the possible units as a table-like structure.

Each column contains all synthesis units which are segmentally equal to the target unit. From this table-like structure a graph is constructed by adding edges between subsequent units. Each path between entries in the first and last column describes one possibility to synthesize the target utterance.

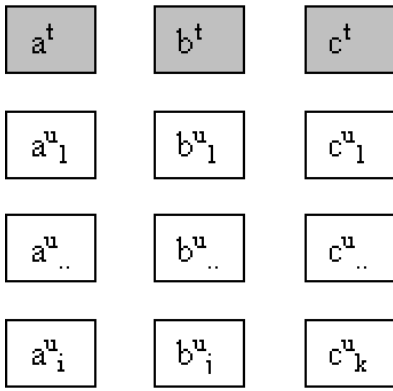


Figure 1: Table-like structure of synthesis units as used for cost based unit selection

If a natural utterance is given as a target then a spectrally motivated similarity between target and synthesis units could be used to find an adequate sequence. It follows that the sum of the unit-based spectral distance minima also gives a good approximation of the minimized spectral distance of the whole utterance. That way, we are spared to regard all possible paths within the graph. The most similar synthetic utterance is achieved by simply sorting the table columns according to the computed similarity between target and synthesis unit. The concatenation of the first row should give the best sequence of synthesis units. In order to determine the second-best synthetic utterance, again all paths describing the correct segmental sequence of units would have to be taken into account. This can be computed efficiently by an extension of the shortest-path algorithm known as n -best-match.

The real problem lies in a useful definition of the spectral distance measure. This measure ought to be motivated perceptually. There are several publications presenting such distance measures. A weighted measure based on the averaged MFCC, power and F_0 coefficients was used as a distance measure and has been applied for clustering units in a synthesis corpus [4]. An evaluation of the correspondence between MOS scores for synthetic and natural speech is reported in [5]. They compare several objective distance measures and their correlation with the MOS scores of a human listening test. In their study, the bispectrum achieves the best results. Since the synthetic utterances already have adequate duration and intonation, this measure seems not to be useful for our work. Quite a different result was reported by [6] where

the MOS values of a listening test correlate highly with a distance measure based on MFCC together with the Euclidean distance. However, all these measures have problems modeling effects known from psychoacoustics. It has been shown that PEMO is well suited for measuring speech quality in speech transmission systems [3]. For this reason, it was used as a pre-processing stage in our approach.

The first step in applying our distance measure is done by representing the target and each synthesis unit by a sequence of 19-dimensional vectors of PEMO coefficients. The vectors are computed at a time shift of 5 ms. As in [4], we will not only compare the vectors belonging to the units, but also 20% of the vectors belonging to each surrounding unit. Experience has shown that this will reduce spectral discontinuities at the concatenation boundaries. A purely PEMO based distance led to unsatisfactory results due to F_0 mismatches. Comparing the mean F_0 between selected and target unit shows an average deviation of 23 Hz. To avoid these mismatches, we extend our measure by a term used for the minimization of F_0 deviations. A second-order polynomial term was used for weighting the F_0 deviation. Small deviations should produce few cost in the distance measure. With increasing deviations the F_0 deviation cost should grow as well. Instead of weighting the F_0 term in (1) by a scalar, we measured the average distance computed by the DTW term. The F_0 deviation cost should be smaller than 1, if the deviation is smaller than 10 Hz. If the deviation is 20 Hz then the cost of the F_0 term should be as high as the average DTW cost. These assumptions are used to determine the unknown quantities of the polynomial (1). Using a second-order polynomial for weighting the F_0 deviation is one possibility - but probably each function of type e^x will be also applicable. Using the additional F_0 term reduces the average F_0 deviations from 23 Hz to 3 Hz.

$$d(x^t, x^u) = \text{DTW}(P^{x^t}, P^{x^u}) + \max \left[0.83 \left| F_0^{x^t} - F_0^{x^u} \right|^2 - 8.2 \left| F_0^{x^t} - F_0^{x^u} \right|, 0 \right] \quad (1)$$

The DTW [7] term in (1) measures the spectral deviation between the sequence of PEMO vectors P belonging to the target unit x^t and a possible synthesis unit x^u .

It is well known that speech prosody is not only made up by F_0 but also by the spectral envelope and segmental duration. Spectral envelope but not phoneme durations are implicitly measured by the DTW term. There are at least two possibilities how duration could be integrated in (1). The first is adding an additional term measuring duration deviation. The second is varying the weight applied to diagonal steps in DTW. Average duration deviation is 23 ms without usage of the F_0 term. Reducing the diagonal weight in DTW from 2 to 1 leads to an average duration deviation of 14 ms. Whether the spectral quality is influenced by this variation needs to be determined via additional listening tests. It is interesting to see that duration deviation with the original definition of diagonal weight [7] decreases to 16 ms if the F_0 term is used. This suggests that duration and F_0 are not independent from each other. Because the application of the F_0 term results in tolerable duration deviation and the correlation between duration deviation and similarity judgments of subjects (cf. Section 4) can be ignored, no further attention was paid to duration deviation in the distance measure.

3. Smoothing at concatenation boundaries

Spectral and phase mismatches, energy discontinuities etc. spoil the perceived quality of the synthetic signal. The manual correction of segment boundaries in large speech corpora is an enormous task. Even if the boundaries are corrected manually, we cannot be sure that discontinuities will not occur during concatenation. Unit selection based synthesis tries to minimize these distortions with the help of continuity distortion terms. Obviously, this will influence the selection of a unit relative to the requested properties because a compromise has to be made between the optimal unit and the discontinuity caused by its use in the synthetic utterance.

If it were possible to avoid discontinuities with the help of a smoothing algorithm, unit selection research could concentrate on the unit distortion terms. In consequence, the computational complexity of the unit selection could be reduced and the search for adequate cost terms could probably be simplified.

Therefore it was decided to perform a smoothing at concatenation boundaries. The smoothing algorithm was motivated by a signal analysis near the boundaries. In particular, low-frequency signal components cause audible disturbances.

These disturbances were reduced with the help of a quadrature-mirror filter (QMF) bank. Such a filter bank divides the signal into subband signals where each subband is spaced of the basis of powers of 2. The original signal can be perfectly reconstructed from the subband signals by applying the inverse filter bank.

Each stage of a QMF filter bank splits the input signal into two signals, say h and g . If n is the Nyquist frequency then h is the band-limited signal which has the frequency range $[0, n/2]$ and g has $[n/2, n]$. The sampling rate of these signals will be reduced by the factor 2. After that, h is reused as input signal for the next filter.

The subbands 250-500, 125-250 and 62.5-125 Hz are weighted by a polynomial of order 2 (parabola). The definition range of this parabola was limited to a fixed size window. Outside this range the signal is not manipulated. The minimum of each parabola lies at the concatenation boundary. The window size was chosen according to the frequency range and sampling rate of the regarded sub-signal. Depending on the window size the parabola reaches 1 at the window boundaries. The window sizes and the parabolas' minima were determined experimentally. It was found that the window size should grow but the parabola minimum should decrease with lower-frequency subbands. This weighting mechanism is applied to both subband signals h and g . The decrease of the parabola steepness was chosen earlier on the h signals. After weighting, the signal was reconstructed. The filter coefficients were calculated using cubic spline approximation as described by [9].

The smoothing technique acts like a time-variant filter which is applied multiply to the subbands. This operation is very similar to a pre-emphasis, the strength of which depends on the distance from the concatenation boundary and on the amplitude of the disturbance.

Informal listening tests performed in our group (7 subjects, 20 utterances) showed a clear preference for the smoothed signals (Table 1). The application of the smoothing algorithm to natural utterances did not lead to any perceptual difference.

	Naturalness	Intelligibility
Raw Concatenation	38 %	37 %
Smoothed Concatenation	72 %	73 %

Table 1: Listeners' preferences for synthetic signals chosen by the distance measure (1) presented both with and without smoothing.

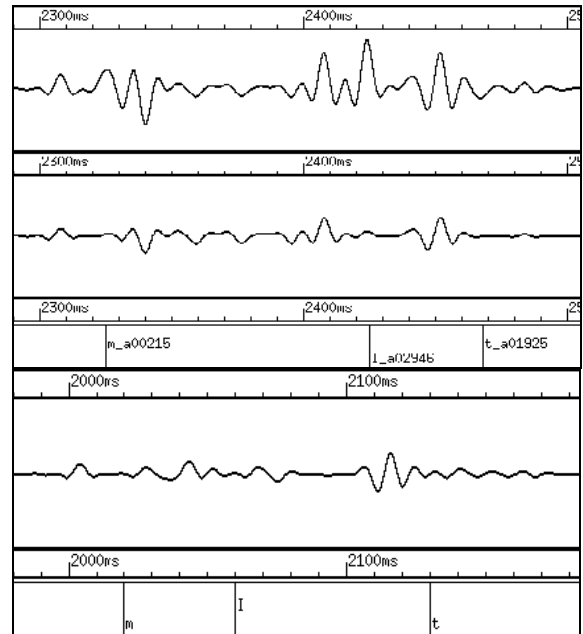


Figure 2: Part of the band-limited signals 62.5-125 Hz. Beginning from the top: raw and smoothed concatenation of the synthetic signal and its segment boundaries; natural signal and its segment boundaries

The presented smoothing method needs further investigation to adapt the smoothing characteristics to the strength of the distortion. Sometimes, the smoothing appears to be too strong, at other times too weak. It should be noted that the smoothing technique is specifically developed for our purpose and unsuitable for smoothing high frequency distortion (e.g. formants).

4. Perception Experiments

Validating the objective distance measure consists of two experiments. The first experiment is used to show that the sorting property of the objective measure coincides with the listeners' judgments (Section 4.1). The second experiment is used to answer the question whether there are alternative unit sequences which are judged more natural but are less similar than those ones predicted by the objective measure (Section 4.2).

The natural utterances used as references and the synthesis units were taken from a large speech corpus which was originally recorded for the Verbmobil speech synthesis [8]. The corpus contains 3.5 hours of speech and consists of approximately 200,000 phonemes. It was prohibited to use units contained in the reference utterance for the construction of the synthetic utterance. This strategy is equivalent to the Leave-One-Out method which is well known from pattern recognition.

4.1. Similarity judgments

It was tested whether the similarity measure results in a ranking comparable to the similarity judgments by human listeners. Therefore phoneme-based units were compared to parts of a natural utterance using (1).

Three utterances with decreasing objective similarity were chosen as synthetic stimuli for the listening tests. Our goal was to show that the distance measure ranks similar to judgments made by humans. Of course, human perception cannot distinguish the same fine-grained acoustic differences as an objective measure. This is for example shown by psycho-acoustically based audio compression (e.g. MP3). Therefore, it was decided to create the synthetic stimuli by concatenating the units of the first three table rows after an application of (1). Usage of the three best alternatives would have resulted in mostly indistinguishable stimuli.

For each of 69 natural utterances, subjects were asked to choose the most similar utterance out of the set of three by comparing it with the natural one. If our algorithm mirrors the human similarity impression, the subjects' choices ought to reflect the preferences of the algorithm. Thus, the initial hypothesis is that the subjects prefer those stimuli categorized most similar by (1).

4.1.1. Experimental Design

Eight phonetic experts participated in the experiment. Subjects listened to the stimuli by clicking on the buttons of a simple GUI. The order of the buttons on the GUI did not reflect the order of distances calculated between each stimulus and its natural counterpart. Instead, the order of stimuli presentation was randomized for each set. This was considered necessary to prevent subjects from searching patterns like "the left button is always the most similar". Subjects were allowed to listen to each stimulus via headphones as often as they wanted to (including the natural utterance).

4.1.2. Results

The results (see Figure 3, Figure 4) clearly indicate that the preferences of the listeners reflect the ranking of the algorithm. The most similar candidate chosen by the algorithm is the most frequent choice of the listeners (χ^2 , $p < 0.0001$, $N = 533$). This result is stable for all subjects (see Figure 4). The stimuli predicted as the most similar ones by the algorithm were chosen by the listeners in almost 50% of all cases. However, the subjects did vary highly among each other with respect to their decisions, and subsequently, their judgments correlate only slightly (Spearman-Rho, $r = 0.2$). It is interesting to see that subjects' agreement is high if subjects prefer the candidate ranked most similar by the algorithm. The disagreement increases if subjects tend to prefer a candidate ranked less similar by the algorithm. This could mean that the synthetic stimuli are too similar thus leading to random judgments. It is also possible that in those cases the distortions induced by unit concatenation influence the subjects' perception too strongly in order to distinguish reliably between the stimuli. Furthermore, no correlation between the subjects' judgment and both F_0 and duration deviation was found.

These results indicate two things:

- Generally, it can be said that the ranking order of all listeners reflects the preferences of the algorithm.
- The more the listeners disagree with the algorithm, the more they also disagree among each other. Apparently, such stimuli are simply difficult to classify.

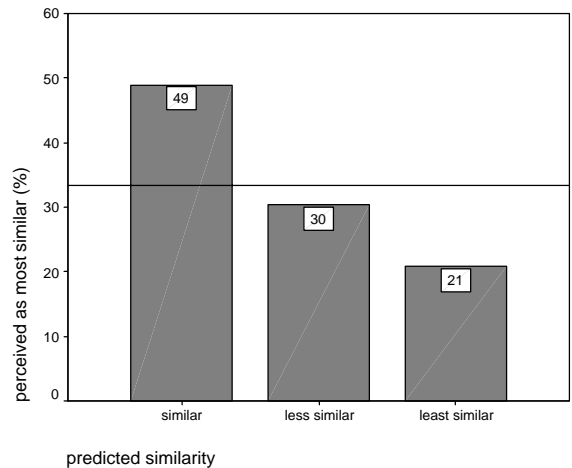


Figure 3: Stimuli perceived as most similar

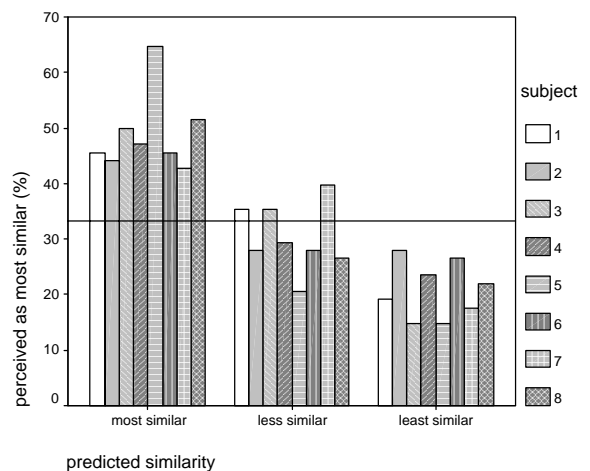


Figure 4: Choices of different subjects

4.2. Similarity versus naturalness

After having shown that our similarity measure is supported by listeners' judgments, the following experiment ought to shed light to the question whether similarity correlates with perceived naturalness, or whether naturalness can be judged independently of similarity relative to a given natural utterance. Therefore it was tested whether the corpus provides us with alternative synthetic variants which are judged more natural than the best similarity-based choice of the distance measure. Again, (1) was applied (with minor variations), but now to differently sized units. We start at the word level and compare word-sized units. If there is no word available in the synthesis corpus we switch to the syllable level. Again, if no syllable is available in the corpus, phoneme-based units are concatenated. We hoped to find out that the spectral smoothness of larger units (which is inherently the best we can achieve) leads

to better naturalness judgments. Since it is often believed that an increase in naturalness can be reached by imitating natural speech as well as possible, a high degree of similarity ought to correlate with high perceived naturalness. Currently, we assume that similarity and perceived naturalness have to be regarded as separate dimensions within unit selection based speech synthesis.

4.2.1. Experimental Design

Ten phonetic experts participated in the experiment. Subjects were presented the natural utterance, the utterance predicted as most similar, and the utterance minimizing the number of concatenation points. As in the first experiment, subjects were allowed to listen to each stimulus as often as they wanted or needed to. For each pair of stimuli, they had to decide for the most similar and the most natural one. Both decisions could be made independently from each other.

4.2.2. Results

The results indicate a clear preference for choosing the utterance created by phoneme-sized units as the most similar one. However, naturalness was perceived better for utterances created of variable sized units. This implies that the smoothing algorithm as well as fine-grained spectral discontinuities may still influence the decision.

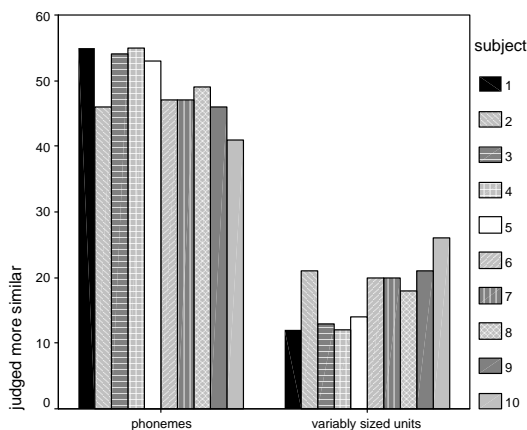


Figure 5: Detailed similarity judgments of different subjects

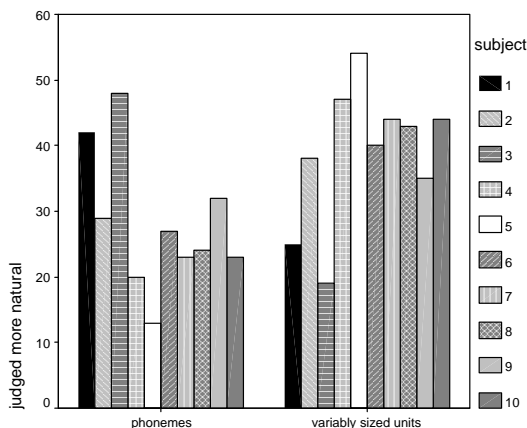


Figure 6: Detailed naturalness judgments of different subjects

	More similar	More natural
phonemes	72%	43%
variably sized units	28%	57%

Table 2: Listeners’ preferences between variably sized units and phoneme-sized units. Variably sized selection give more natural, phoneme-sized selection gives more similar results.

5. Discussion

It has been shown that the spectrally motivated distance measure is useful for predicting units which ought to be selected by the unit selection algorithm because they minimize unit distortion (Section 4.1). It has also been shown that discontinuity distortion can currently not be entirely substituted by a smoothing algorithm (Section 4.2). Minimizing continuity distortion implicitly leads to a preference of unit sequences produced consecutively in the carrier sentences. Section 4.2 shows that this strategy will not *always* lead to the most natural result. If the goal of unit selection is to find a sequence of units which on average is most similar to a given natural utterance, then continuity distortion is an impediment. It could be concluded from Section 4.2 that unit selection should concentrate in future on minimizing unit distortion and that continuity distortions should be minimized by some spectral smoothing procedure. In our opinion, signal manipulation – including small prosodic modifications – would be a necessity for high quality unit selection based synthesis.

“Naturalness” is a term describing a large set of listener impressions. Many of the dimensions contained in this set depend on the world knowledge of the listener and are currently inexplicable. We believe that for each utterance there exists a region within a “naturalness space”. This region has to be defined on the basis of listeners’ agreements. An utterance must be natural if it is positioned within this naturalness space. Each natural utterance must lie within this space.

Similarity between a natural and a synthetic utterance can be measured using techniques available today. Perfect similarity (this should not mean that the signals are mathematically identical) between synthetic and natural signal implies natural synthesis. This means that sufficient similarity in a large number of cases necessarily leads us to naturalness.

It should be kept in mind that listeners’ similarity preferences correlate poorly. We are uncertain how this can be explained. The experiment was designed as simple as possible. Therefore, only a binary decision was asked for. Several subjects reported problems with the task because the utterances were often either too similar or spectral discontinuities disturbed their perception. Future listening tests should pay attention to this problem and enable the listeners to rate two stimuli as perceptually equal.

6. Conclusions

We conclude that using (1) is a valid method to determine units which should be selected from the unit distortion terms in a unit selection based synthesis. However, (1) is only a first step halfway towards a natural-sounding synthesis. Further development of objective distance measures should concentrate on the short-time spectral discontinuities. Therefore, the

currently used temporal integration of locally measured distances cannot provide an appropriate solution. At a first glance, the first derivation of the sequence of parameter vectors could help us finding short-time spectral discontinuities. However, such discontinuities are also properties occurring in natural speech (e.g. plosives), and it is a problem to differentiate among the different types of distortion.

Unlike the results reported by [10], spectral smoothing improves the perceptual quality compared to raw concatenation (Table 2). Smoothing might substitute the continuity distortion terms which make up a major factor of the unit selection complexity. Pre-selecting larger units provides an alternative way to do unit selection without continuity distortion terms and leads to increased naturalness. But we still need a rejection criterion telling us when a smaller unit ought to be preferred, thus sacrificing spectral smoothness for a higher degree of similarity. Measuring the similarity between one large unit and the best sequence of smaller units substituting this larger unit may be the right way towards reaching this goal. Further work will integrate such a criterion into our unit selection algorithm.

7. References

- [1] Black, A.W., Campbell, N., (1995), "Optimising selection of units from speech databases for concatenative synthesis", Proc. of the Eurospeech'95, Vol. 1, pp. 581-584.
- [2] Sonntag, G., (1999), "Evaluation von Prosodie", Doctoral Thesis, University of Bonn, Shaker, Aachen.
- [3] Hansen, M., (1998) "Assessment and prediction of speech transmission quality with an auditory processing model", Doctoral Thesis, University of Oldenburg.
- [4] Black, A.W., Taylor, P., (1997), "Automatically clustering similar units for unit selection in speech synthesis", Proc. of the Eurospeech'97, Vol. 2, pp. 601-604, Rhodes, Greece.
- [5] Chen, J.D., Campbell, N., (1999), "Objective distance measures for assessing concatenative speech synthesis", Proc. of the Eurospeech'99, Vol. 2, pp. 611-614, Budapest, Hungary.
- [6] Wouters, J., Macon, M.W., (1998), "A perceptual evaluation of the distance measures for concatenative speech synthesis", Proc of ICSLP '98, Paper number 905, Sydney, Australia.
- [7] Sakoe, H., (1978), "Dynamic programming algorithm optimisation for spoken word recognition", IEEE Trans. ASSP, Vol. 26, pp 43-49.
- [8] Stöber, K., Wagner, P., Helbig, J., Köster, S., Stall, D., Thomae, M., Blauert, J., Hess, W., Hoffmann, R., Mangold, H., "Speech Synthesis by Multilevel Selection and Concatenation of Units from Large Speech Corpora", In Wolfgang Wahlster (editor), "Verbmobil: Foundations of Speech-to-Speech Translation", Symbolic Computation, Springer, (Berlin, 2000), pp. 519-537.
- [9] Mallat, S., (1989) "A theory for multiresolution signal decomposition : the wavelet representation", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 11, p. 674-693, July 1989.
- [10] Chappell, D. T., Hansen, J. H. L., (1998), "Spectral Smoothing for Concatenative Speech Synthesis", Proc of ICSLP '98, Vol. 5, pp. 1935-1938, Sydney, Australia.