Reducing spectral mismatches in concatenative speech synthesis via systematic database enrichment.

Maria Founda, George Tambouratzis, Aimilios Chalamandaris, George Carayannis

Institute for Language and Speech Processing
6, Artemidos str. & Epidavrou, Paradissos Amaroussiou 151 25, Athens, Greece
Email: {mfounda, achalam, giorg_t, gcara}@ilsp.gr

Abstract

This paper presents work performed for the Time-Domain TTS system, which is being developed at the ILSP for the Greek language. It focuses on the enhancement of the synthetic speech quality, by reducing the spectral mismatches between concatenated segments. To that end, a study has been performed to determine the distance that can best predict when a spectral mismatch is audible. Experimentation with different spectral distances has taken place and the distance with the best performance has been used in order to systematically enrich the segment database, which initially contained only one instance per segment. Results of this procedure indicate a substantial improvement on the synthetic speech quality.

1. Introduction

This work focuses on research associated with the new ILSP Text-To-Speech (TTS) system, which is based on the concatenative synthesis paradigm, speech being synthesized by concatenating diphones. In such systems the generated synthetic speech although highly intelligible, tends to sound unnatural. This is mainly attributable to the mismatches that take place at the joints of the diphones and to the distortion injected by the prosodic modifications of the speech signals. In order to deal with this problem, the development of a sophisticated algorithm that chooses the best matching unit is required [1], [2]. This solution requires that the database contains several instances of each diphone taken from different contexts, and therefore with different prosodic and spectral characteristics. During synthesis, the most appropriate instance should be chosen, so that both the mismatches at the joints and the prosodic modifications required are minimized.

The goal of this paper is to study a specific source of distortion, namely the spectral mismatch between the concatenated diphones. In our case, instead of performing spectral manipulation to smoothen the joints at unit boundaries, spectral discontinuities are reduced by providing multiple diphone instances in the database. Then, a unit selection algorithm is employed to choose the best matching diphone-sequence.

In order to achieve this, four spectral distances were compared in an attempt to find an optimal measure of audible spectral discontinuities. This step was followed by a systematic enriching procedure of the initial database (which contained only one distance of each diphone), in an attempt to minimise the mismatches at joints between vowels. The enriching was initially restricted to vowels because it has been reported that spectral mismatch at diphone joints has its greatest effect within vowels [2], [3]. Finally the quality of the synthetic speech generated by the enriched database was evaluated via both objective (analytical) and subjective methods.

2. Experimenting with four spectral distances

In this section, the effect of spectral mismatches on synthetic speech was examined via specifically designed experiments. The focus was to determine the distance that most accurately indicates when a spectral mismatch is audible. To achieve that, listening tests were performed, using recorded speech samples.

2.1 Description of listening tests

To implement the listening tests, initially speech material was obtained from a trained male speaker. The test utterances used in the experiment were presented as sets of phrases called test sets. In each test set the same phrase was repeated, having each time a single diphone replaced with other marked instances of the same phonetic identity originating from different contexts. All prosodic characteristics (duration, power and pitch contour) were smoothened to the values of the original diphone, thus leaving only one major source of distortion, the spectral mismatch.

A combination of labeled samples and natural utterances provided reference points throughout the experiment to ensure the accuracy of the experimental results. A total of 31 subjects were used as listeners. For all subjects high-quality equipment were used as listeners. For all subjects high-quality equipment was employed, in order to make even the slightest distortion audible (a detailed description of this procedure is given in [4]).

2.2 Evaluation of the listeners

The set of listeners can be divided into two groups: Group1 consisting of members who had a background in speech processing, and Group2 consisting of members who had no such background. Group1 consisted of 11 listeners and Group2 of 20 listeners.

The marks attributed to the speech samples by the listeners were evaluated in terms of accuracy and consistency. This was performed in two ways:

- In each test set to be evaluated by the listeners, the natural utterance was included. Listeners who repeatedly marked these utterances with low scores should be excluded from the subset used to evaluate the spectral distances.
- In addition, a certain test set was presented both at the beginning and at the end of each listening test. This was used to reveal whether each listener was consistent in the
way he evaluated the speech samples during the entire experiment.
Using this information, six listeners were excluded. Of these, one belonged to Group1 and the remaining five to Group2. Subjects of Group1 seemed as a whole to be more consistent in the way they evaluated the natural speech samples, as would be expected due to their background.

2.3 Spectral distances
Four distances were used as candidates for expressing spectral mismatches. The Euclidean Formant Distance, $E_F$ ([3]), is given by the following formula:

$$E_F(F, F') = \sqrt{\sum (F_i - F_i')^2}$$  \hspace{1cm} (1)

where $F_i$ and $F_i'$ are the $i$-th formants at the respective boundaries of the two diphones where the formant mismatch is to be calculated.

The Euclidean Distance of the formants’ magnitude $E_a$ is calculated by:

$$E_a(a, a') = \sqrt{\sum (a_i - a_i')^2}$$  \hspace{1cm} (2)

where $a_i$ and $a_i'$ are the magnitudes of the formants $F_i$ and $F_i'$, respectively.

The Weighted Euclidean Formant Distance $E_{wF}$ is defined as:

$$E_{wF}(F, F') = \sqrt{\sum w_i (F_i - F_i')^2}$$  \hspace{1cm} (3)

where the weights are functions of the magnitudes of the formants. The formants of the three preceding distances were estimated using peak-picking on the LPC-spectral envelopes. The spectral envelopes were estimated for the frequency range of 0 to 5000 Hz.

Finally, the Kullback-Leibler Distance $KL$ ([3]) is defined by the following formula:

$$KL = \int f(x) \cdot \log \left( \frac{f(x)}{g(x)} \right) dx$$  \hspace{1cm} (4)

The latter distance was calculated via the two power-normalized spectral envelopes, estimated using Hanning windows positioned at the boundaries of the two diphones.

2.4 Statistical analysis of the experimental results
To investigate the effectiveness of the different distances, a series of 7 listening test sets was studied in more detail, where the replaced diphone included at least one vowel or voiced consonant. The distances were calculated using formulae (1) to (4), and summing for the left and right discontinuities. These distances were compared to the average score from Group1 and the average score from Group2, using a regression analysis. It was found that:

- for five test sets, the KL distance provided the most accurate prediction of the subjective quality of the synthetic speech, which was statistically significant at a level of 5%;
- for one of the five aforementioned test sets, the $E_F$ distance gave results of an equivalent quality to the KL distance, which were also statistically significant at a level of 5%;
- for the remaining two test sets, none of the four studied distances gave statistically significant results at a level of 5%. In one case, $E_F$ gave the best results while in the other one the KL distance gave the best results (which were significant only at a level of 10%).

To summarize, the KL distance was found to be the most accurate predictor of synthetic speech quality (and thus of spectral mismatches) and was also the only distance to consistently generate statistically significant results. The best results were obtained for diphones including vowels, indicating that using these distances this phonemic category is probably more amenable to improvements in the synthetic speech quality by introducing new instances.

3. Database vowel enriching using the Kullback-Leibler distance
As noted earlier the initial database contained only one instance of each diphone, extracted from a given context. Thus, when this diphone is used in a different context during synthesis, an audible mismatch is likely to occur. The aim is to enrich this database in an attempt to minimize the spectral mismatches that take place in the middle of a synthesized vowel. For this purpose, the Kullback-Leibler (KL) distance was used. However the procedure of database enrichment is a time-consuming task, as it requires the introduction of additional instances of diphones. Moreover, the size of the database needs to be limited to manageable proportions. For that reason it was decided to follow a systematic approach in which multiple instances in the database are provided only for diphones causing large spectral mismatches.

3.1 Determining the diphones for which multiple instances are required
In the ILSP TTS system, distinct phonemes are used for stressed and unstressed vowels, resulting in 10 vowels for the Greek language. The distribution of the KL distances for all the joints within each of the ten vowel phonemes was calculated, that is for diphone pairs (/X1-v/, /v-X2/). In this case, $v$ stands for the semi-vowel being studied, while $X_1$ and $X_2$ range over all half-phones of the Greek language. A typical diagram, referring to vowel /a/, is shown in Figure 1 where the values of the KL distances are normalized to the range [0,1].

Distributions such as that of Figure 1 serve to indicate the number of additional diphones that need to be inserted in the database in an attempt to eliminate mismatches greater than a certain threshold. To determine the diphones for which alternative instances should be inserted (the insertion process being referred to as re-marking) an inventory of pairs of diphones is created, referred to as List1. List1 contains all pairs of diphones whose concatenation gives a KL distance that is greater than the threshold value. Distributions similar to the one of Figure 1 also provide an indication of the spectral mismatches that take place when constructing one vowel by concatenating the diphones from the database.

It should be noted that the maximum value of the KL distance was not the same for all vowels. Stressed vowels
were found to have smaller spectral mismatches than unstressed ones, at least for frequencies up to 5000Hz.

There are two reasons why this inventory of diphone-pairs is not sufficient when determining which diphones require multiple instances:

- It is possible that some diphones form one (or very few) large distances, while other diphones have in general substantial values but none of them exceeds the threshold value.
- It could also be that only one of the two diphones of the pair that gives a high distance is ‘problematic’ (i.e. it generates large spectral mismatches when concatenated with other diphones).

For that reason a second metric has been obtained for each vowel. This is the mean value (as well as the variance) of the KL distances of each diphone of the form /X1-v/ with all the diphones of the form /v-X2/, that can follow diphone /X1-v/ when speech is synthesized. The same mean values are computed for all the diphones of the form /v-X2/ for the particular vowel. A typical example (corresponding to vowel /a/) is shown in Figure 2.

Following these calculations, the diphones that form in general large spectral mismatches when combined with the other diphones of the database are detected. The diphones that need to be re-marked are those whose mean distance value is high (irrespective of the variance value) as well as those whose variance is high. Diphones from this second category are included since a high value in the variance indicates a large number of values that deviate considerably from the mean value.

This way an inventory of diphones to be re-marked is created, referred to as List2. List2 is then compared with List1, ensuring that at least one diphone from each pair of List1 is also included in List2. Following this process, List2 provides the final list of diphones to be re-marked.

3.2 The process of re-marking different instances of diphones

For each diphone of List2, two extra instances were re-marked. These were selected from a different context each time, so that the vowel of the diphone has as context a fricative, an explosive and another vowel, nasal or liquid. About 30-40 instances were re-marked for each vowel as it was found that this amount provides a substantially improved quality in terms of spectral mismatch, as speech is synthesized.

4. Evaluation of the enriched database

To evaluate the results of this enriching procedure, the spectral mismatches of all the new pairs of the form /X1-v/, /v-X2/, were compared to the spectral mismatches of the respective combinations in the initial database. The comparison was performed using both a theoretical and a practical approach, as described in the following section.

4.1 Objective evaluation

The additional instances of the new database resulted in groups of pairs that had the same phonetic content but were formed from different diphone instances, thus having different spectral mismatches at the joints. One simple example is given in Table 1 for the combination of diphones /a’e/ and /ee/, where the first row corresponds to the instance existing in the original database. The improvement in the KL distances following the introduction of alternative instances is evident.

```
Table 1: set of instances for the diphone combination /a’e/+/ee/, together with the corresponding KL distances.

<table>
<thead>
<tr>
<th>Diph. 1 identity</th>
<th>Diph. 2 identity</th>
<th>KL distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1071</td>
<td>432</td>
<td>0.7534</td>
</tr>
<tr>
<td>1071</td>
<td>1221</td>
<td>0.5937</td>
</tr>
<tr>
<td>1223</td>
<td>432</td>
<td>0.2118</td>
</tr>
<tr>
<td>1223</td>
<td>1221</td>
<td>0.4222</td>
</tr>
</tbody>
</table>
```

For such groups with multiple instances, the minimum, mean and maximum values of the KL distance of each group are calculated and compared to the respective (unique) value of the initial database. The results of such a comparison for vowel /e/ are shown in Figure 3. Values below the diagonal are smaller than the respective initial ones, indicating a
reduction in spectral mismatch magnitudes in comparison to the original dataset.

Figure 3: the minimum, mean and maximum value of the KL distance, for each group of diphone pairs forming the vowel /e'/, in relation to the initial (unique for each group) values.

Figure 4: distributions of the minimum, mean and initial KL distances, as compared to those of the original database (which are also plotted in the Figure).

Figures 3 and 4 indicate that the smallest values achieved for each group (such as the one depicted in Table 2) are significantly lower than the initial ones. Since spectral mismatch is not the only factor considered during the unit selection process, the unit selection algorithm shall not always choose the combinations with the lowest spectral mismatches. Nevertheless, since the mean KL distance value of each group remains also quite low (between 0.1 and 0.3), in the enriched database there exist several alternative combinations with relatively small spectral mismatches. One of these should fulfill the criteria of the unit-selection algorithm and thus will most likely be chosen. It should also be noted that even when the initial spectral mismatches are large, the respective least and mean values in the new database remain consistently low. For example, as depicted in Figure 3, in the case of vowel /e'/ distances greater than 0.3 (normalized KL distance) would most probably be avoided. The results obtained for the rest of the Greek vowels were of a similar quality, indicating the effectiveness of this systematic method when enriching the diphone database.

4.2 Subjective evaluation

In order to examine the improvement in the synthetic speech, an acoustic experiment was performed using the enriched database together with an optimal unit selection algorithm. The material used consisted of 18 words of the Greek language, containing several instances of vowels. These words were synthesized using both the initial database and the enriched database in which diphones were selected using the algorithm. Five subjects were asked to judge which of the two instances of each word sounded better. The results are summarised in Table 2, where it is seen that words synthesized using the enriched database get a consistently higher score in 70% of the cases.

Table 2: Collective results for 18 synthetic utterances.

<table>
<thead>
<tr>
<th>Preferred utterance</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched database</td>
<td>70 %</td>
</tr>
<tr>
<td>Initial database</td>
<td>18 %</td>
</tr>
<tr>
<td>Equivalent quality</td>
<td>12 %</td>
</tr>
</tbody>
</table>

5. Conclusions

A method for enhancing the speech output quality of concatenative TTS systems has been presented in this paper. This is based on systematic enriching of the initial database so that large spectral mismatches within vowels are eliminated. The method has been shown to result in substantial reduction of the spectral mismatch magnitude in the synthetic speech. In the future more diphones will be added in the database, in order to further reduce the mismatches across diphone boundaries. In addition, voiced consonants are to be studied.

6. References