A SOLUTION TO THE REDUCTION OF CONCATENATION ARTEFACTS IN SPEECH SYNTHESIS

Esther Klabbers, Raymond Veldhuis and Kim Koppen

IPO, Center for User-System Interaction, Eindhoven, The Netherlands
{E.A.M.Klabbers/R.N.J.Veldhuis}@tue.nl

ABSTRACT

One problem with speech synthesis impedes high quality is the occurrence of audible discontinuities at segment boundaries. Formant jumps across concatenation points suggest the problem to be due to spectral differences. The problem is most apparent in vowels and semi-vowels. We propose to reduce the number of audible discontinuities by adding context-sensitive diphones to the database. The number of additional diphones is limited by clustering contexts with similar spectral effects on the neighbouring vowels, using the Kullback-Leibler distance. A listening experiment has shown that the percentage of perceived discontinuities has significantly decreased.

1. INTRODUCTION

One problem with diphone concatenation impedes high quality is the occurrence of audible discontinuities at diphone boundaries. Discontinuities are caused by mismatched diphons in F0, phase or spectral envelopes across concatenation points [2]. In the IPO speech engine Calipso, F0 mismatches are avoided by monotising the diphones before storing them in the database. Phase mismatches are avoided by using a method called phase synthesis for resynthesis of the nonsense words [3]. The analysis is pitch-synchronous, using overlap-and-add over two pitch periods. Because the harmonic components of adjacent pitch periods are added with coherent phases, phase mismatches are avoided.

Spectral mismatch is still a major problem, though. In a previous experiment [4], the results of a listening experiment concerning five vowels /a/, /i/, /A/, /I/ and /u/ in CVC-context were related with several spectral distance measures to find a measure that best predicts the audible discontinuities. The Kullback-Leibler (KL) distance, coming from statistics, was found to be the best predictor.

2. SYMMETRICAL KULLBACK-LEIBLER DISTANCE

The Kullback-Leibler (KL) distance or relative entropy is used in statistics [5] to compute the distance between two probability distributions. Here, it is calculated from the two power-normalised spectral envelopes \( P(\omega) \) and \( Q(\omega) \). The original asymmetrical definition of the KL distance is changed into a symmetrical version:

\[
D_{SKL}(P, Q) = \int (P(\omega) - Q(\omega)) \log \left( \frac{P(\omega)}{Q(\omega)} \right) d\omega, \quad (1)
\]

assuring that \( D_{SKL}(P(\omega), Q(\omega)) = D_{SKL}(Q(\omega), P(\omega)) \).

It has the important property that it emphasises differences in spectral regions with high energy more than differences in spectral regions with low energy. Thus, spectral peaks are emphasised more than valleys between the peaks and low frequencies are emphasised more than high frequencies, due to the 6 dB/octave declination in spectral energy of speech signals.

3. CLUSTERING PROCEDURE

In order to reduce the number of audible discontinuities, we propose to extend the diphone database with context-sensitive diphones. We keep the size of the database within bounds by clustering contexts that are spectrally alike according to the SKL distance measure. Suppose we divide the diphone sets \( C_l V, l = 1, \ldots, M \), and \( V C_r, r = 1, \ldots, M \), for a particular vowel \( V \) into two sets of \( N \) clusters \( L(V)_k \) and \( R(V)_k, k = 1, \ldots, N \), with maximum SKL distance across diphone boundaries in corresponding clusters \( L(V)_k \) and \( R(V)_k \) below a threshold \( \beta \). The maximum SKL distance between two non-corresponding clusters \( L(V)_k \) and \( R(V)_l \), with \( k \neq l \), will not be limited to \( \beta \). We now construct additional clusters \( R(V)_{1,k} \), \( k \neq 1 \), containing the diphones of \( R(V)_1 \), but recorded with a left-side context consisting of a representative diphone in \( L(V)_k \). Instead of concatenating a diphone from \( L(V)_k \) with one from \( R(V)_1 \), a diphone from \( R(V)_{1,k} \) will be used. This
hopefully reduces the maximum SKL distance across di-
phone boundaries to a value lower than \( \beta \), although a guar-
antee cannot be given in advance. This procedure will in-
crease the number of VC diphones for a particular vowel
by a factor \( N(\beta) \), which is equal to the number of clusters.

Figure 1 illustrates the clustering procedure for the vow-
el /u/. In our investigation, the maximum number of clus-
ters is restricted to three, which contain the same conson-
antal contexts for left and right diphones. Adding more
than three clusters will not significantly reduce the proba-
bility of an audible discontinuity. Concatenating /bu/ and
/uk/ to make boek is expected to be unproblematic, as both
consonants come from the same cluster with a small SKL
distance. However, for the words doek and hoek an au-
dible discontinuity is likely to occur as they come from
non-corresponding clusters. As a remedy, additional right
diphones are recorded with a representative from a non-
corresponding left cluster. This means that for the /uk/
diphone, which was originally recorded in the symmet-
rical nonsense word k@fuk@, two additional diphones are
recorded in the asymmetrical nonsense words l@fuk@ and
f@fuk@. Then, the word doek can be created by concate-
nating the original left diphone /du/ with the new diphone
/uk/ taken from l@fuk@ and the word hoek is created by
concatenating the same /du/ diphone with the new right di-
phone /uk/ taken from f@fuk@.

The clusters are constructed according to a classifica-
tion algorithm, derived from the LBG (Linde-Buzo-Gray)
algorithm [8]. The SKL distance is used as a criterion
for the division. A Distance Matrix (DM) is constructed
with C_0 V diphones in the rows and VC_r diphones in the
columns. The clustering procedure works as follows:

1. Three C_0 V diphones are chosen as the initial repre-
sentatives of the clusters.
2. The distance matrix is reduced to a cluster matrix
with SKL distances between the three representa-
tives and the VC_r-diphones. Each VC_r-diphone is
added to the cluster to which representative it has
the lowest SKL distance.
3. The initial representative does not necessarily have
the lowest average SKL distance to all other diphones
in the cluster, so for each cluster a new representa-
tive is chosen that does adhere to this criterion. Then
steps 2 and 3 are repeated until the cluster configu-
ration converges.

All combinations of initial representatives were tried.
Those with the lowest maximal distance in a cluster are
displayed in Table 1. There is no clear pattern related to
manner or place of articulation of the consonants, except
for the /u/ where all alveolars end up in the same cluster.
We will come back to this issue in the discussion. The clus-
ter configuration for /u/ was already visualised in Figure 1.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Consonants in cluster</th>
<th>Maximum SKL</th>
<th>Average SKL</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a:/</td>
<td>1: v bfwz</td>
<td>1.08</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>2: S</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>3: x GJLNdghjklmprstz</td>
<td>3.13</td>
<td>0.47</td>
</tr>
<tr>
<td>/i:/</td>
<td>1: k GNgpstkx</td>
<td>2.40</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>2: b JLZjmrwvwz</td>
<td>2.79</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>3: S chl</td>
<td>1.85</td>
<td>0.89</td>
</tr>
<tr>
<td>/u:/</td>
<td>1: G Nbgkvwx</td>
<td>2.36</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>2: JLSZcdjnstz</td>
<td>2.02</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>3: f hmpr</td>
<td>2.39</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 1: Cluster configurations for /a:/, /i:/, and /u/. The representatives in each cluster are the first consonants in each row.

4. PERCEPTUAL EXPERIMENT

In order to measure the improvement that results from the
addition of context-sensitive diphones new recordings were
made with which a perceptual experiment was performed.
In order to make comparison possible, the new recordings
contained both the old symmetrical and the new asymmet-
rical nonsense words.

4.1. Material

Stimuli were created consisting of concatenated C_1 V and
VC_r-diphones. Like in the first experiment, the conso-
nant portions were cut off, to prevent them from influenc-
ing the perception of the diphone transition in the middle
of the vowel. For each C_1 V and VC_r combination there
were two versions, one with a right diphone from the sym-
metrical nonsense word C_0 @ C_r VC_r @ (database without
clustering) and one with a right diphone from the asym-
metrical nonsense word C_0 @ C_0 C_r VC_r @ (database with
clustering). In order to reduce the total number of stimuli,
it was decided to focus on just three vowels /a:/, /i:/, and
/u/. The total number of stimuli used in the experiment
is 2254, of which 1449 were constructed according to the
original concatenation method (21 C_1 V, 295 for
/u/). The total number of diphone combinations were obtained using diphones
from the context-sensitive database (202 for /a:/, 295 for
/i/ and 308 for /u/, based on three clusters).

4.2. Procedure

The perceptual experiment of [4] was repeated, this time
using six participants with a background in psychoacous-
tics or phonetics. They had not taken part in the previ-
ous experiment. The participants had to judge whether the
diphone boundary in the middle of the vowel was either smooth or discontinuous. The stimuli were presented in three hourly sessions which were held on three different days. Each session was split into two 30-minute blocks by a 15-minute break. The session order was different for all participants. The set-up of this experiment results in very critical observations because the vowels have been placed out of context and a binary decision has to be made.

4.3. Results

Table 2 lists the percentage of perceived discontinuities for the new database with and without clustering. As in [4], these are based on the majority scores (4 out of 5 listeners agree). Since we had six participants in this experiment, one randomly chosen subject was left out to keep the results comparable to the old situation. The results for the new database without clustering is better than for the original database. This can be a result of more careful pronunciation on the part of the speaker. Nevertheless, it can be seen that clustering does reduce the number of audible discontinuities.

Its significance is demonstrated by a repeated measures ANOVA which was performed on the SKL distance and on the summed participants’ scores. The results are presented in Table 3. When looking at the results for the SKL distance one can observe that the distance has significantly decreased for context-sensitive diphones for both /i/ and /u/, but is not significant for /a/. However, in the judgement of the participants, the number of detected discontinuities has significantly decreased for all three vowels.

![Figure 1](image)

**Figure 1:** The principle of the construction of additional diphone clusters. The context-sensitive diphone-clusters are indicated in grey. They consist of VC\_diphones that were recorded with a representative from a non-corresponding left cluster. The representative of each cluster is circled.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>New database without clustering</th>
<th>New database with clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>7.7%</td>
<td>6.0%</td>
</tr>
<tr>
<td>/i/</td>
<td>19.3%</td>
<td>17.0%</td>
</tr>
<tr>
<td>/u/</td>
<td>53.4%</td>
<td>26.3%</td>
</tr>
</tbody>
</table>

**Table 2:** Percentage of perceived discontinuities for each vowel. The percentages are computed from the sum of the majority scores.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>SKL distance</th>
<th>Sig</th>
<th>Sum of participants’ scores</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>F_1,482 = 0.27</td>
<td>n.s.</td>
<td>F_1,482 = 7.68</td>
<td>*</td>
</tr>
<tr>
<td>/i/</td>
<td>F_1,482 = 40.02</td>
<td>*</td>
<td>F_1,482 = 6.85</td>
<td>*</td>
</tr>
<tr>
<td>/u/</td>
<td>F_1,482 = 65.95</td>
<td>*</td>
<td>F_1,482 = 155.40</td>
<td>*</td>
</tr>
</tbody>
</table>

**Table 3:** Repeated measures ANOVA on new database with and without clustering for SKL distance and summed participants’ scores; Sig indicates significance (n.s. = not significant, * = significant).
When again considering the specific example of /dük/, it can now be seen that adding an additional /u:k/ diphone that has been recorded in the appropriate context makes a visible difference (compare the top and bottom panel of Figure 2). Instead of an abrupt and large jump in the F2 as observed in the top panel, the F2 descends more gradually.

Figure 2: Improvement in the concatenation of /dük/ and /u:k/ using a different diphone for /u:k/.

5. DISCUSSION AND CONCLUSION

The finding that audible discontinuities still occur for the /a:/ and that clustering does not reduce the amount of audible discontinuities leads to the conclusion that besides coarticulation there is always random variation in the pronunciation of the stimuli. This was also observed by [6] who found F2 variations in excess of 50 Hz for a vowel in repetitions of the exact same phrase as uttered by a highly professional speaker. [7] reports even larger F1 and F2 variations (up to 250 Hz) in the repeated pronunciation of /i/ in six and million by a professional speaker. This indicates the need to record several instances of a nonsense word and choose the one that is optimal for the database.

The bottom panel in Figure 2 shows that when the diphones are recorded in asymmetrical nonsense words, the formant trajectories are no longer stable, but change gradually from start to finish. In that case, it may make sense to optimise the cutting point of the diphone boundary as proposed by [1].

This paper reported on a solution to the reduction of the occurrence of audible discontinuities in diphone synthesis caused by spectral mismatch at the diphone boundaries. Context-sensitive diphones were added to the database. In order to reduce the number of additional diphones, the SKL distance was used to cluster consonantal contexts that have the same spectral effects on the neighbouring vowels. A perceptual experiment was conducted to evaluate the improvement obtained with this addition to the database. A significant improvement was obtained for /a:/ and /i:/ both in terms of the objective SKL distance and the subjective scores. For /a:/ there was only a subjective improvement, but objectively, in terms of SKL distance, the improvement was not significant. This is not a problem, however, as the number of discontinuities in /a:/ was already low to begin with.

Although the research was performed on a restricted type of stimuli, we think the procedure of detecting audible discontinuities using the SKL measure is also applicable to other stimuli. Currently, speech synthesis using on-line selection of variable length units is very popular. We expect that the SKL measure can be successfully integrated in this approach to select the best fitting units.

REFERENCES