ON THE REDUCTION OF CONCATENATION ARTEFACTS IN DIPHONE SYNTHESIS

Esther Klabbers and Raymond Veldhuis*

IPO, Center for Research on User-System Interaction, Eindhoven, The Netherlands
{klabbers/veldhuis}@ipo.tue.nl

ABSTRACT

One well-known problem with diphone concatenation is the occurrence of audible discontinuities at diphone boundaries, which are most prominent in vowels and semi-vowels. Significant formant jumps at certain boundaries suggest that the problem is of a spectral nature. We have examined this hypothesis by correlating the results of a listening experiment with spectral distances measured across diphone boundaries. The aim is to find a spectral distance measure that best predicts when discontinuities are audible in order to find out how the diphone database can best be extended with context-sensitive diphones. The results show that the Kullback-Leibler measure is the best predictor.

1. INTRODUCTION

Most speech synthesis systems available today are based on diphone concatenation. One well-known problem with diphone concatenation is the occurrence of audible discontinuities at diphone boundaries, which are most prominent in vowels and semi-vowels and are caused by contextual influences. Our speech-synthesis system currently uses diphones from a professional female speaker that were recorded embedded in nonsense words. Figure 1 shows the spectrogram for the vowel /u/ in the synthesized Dutch word /zuk/. It reveals a considerable jump in F_2 of around 500 Hz at the diphone boundary (between 180 and 185 ms). This, together with other informal observations, suggests that the problem is of a spectral nature. Other causes are discussed in [2]. Several approaches have been proposed to solve this problem:

- The number of audible discontinuities can be reduced by using larger units such as triphones. This does not solve the problem as discontinuities continue to occur albeit less frequently. Moreover, the inventory size increases drastically.
- Spectral mismatch can be minimized by varying the location of the diphone boundary dependent on the context [1]. This calls for a spectral distance measure that correctly represents the amount of spectral mismatch. Moreover, it is based on the underlying assumption that the formant trajectories are not flat, such that shifting the cut-point reduces the size of the formant jump. Figure 1, along with many other observations in our database, shows that formant trajectories can be fairly flat throughout a vowel. The F_2 of the /u/ in /zu/ remains at around 1500 Hz while the F_2 of the /u/ in /uk/ decreases very slowly from approximately 1000 Hz to 800 Hz.
- Spectral mismatch can be reduced by wave-form interpolation, spectral-envelope interpolation or formant trajectory smoothing, the latter of which is preferred [2]. It requires a signal representation that allows this type of operation. The disadvantage of formants as a representation is that they are very difficult to estimate reliably. Wave-form and spectral envelope interpolation have the disadvantage that smooth transitions are often achieved at the expense of naturalness.
- The number of discontinuities can be reduced by including context-sensitive or specialized units in the database [8]. This implies that one knows which contexts can be clustered so as to keep the inventory size within bounds. Our investigation is aimed at gaining insight in this approach.

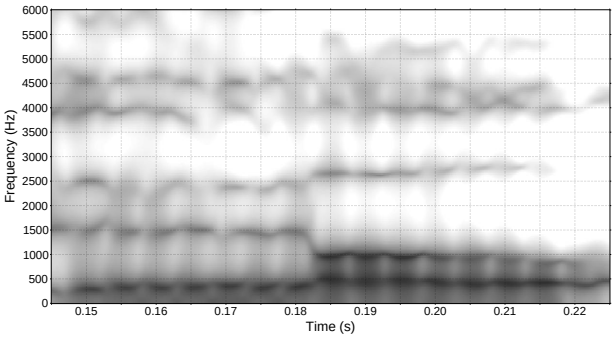

Figure 1: Spectrogram showing spectral mismatch in /u/

We present a detailed study of the occurrence of audible discontinuities in our diphone database. To reduce the data set to manageable proportions, we restricted the study to five Dutch vowels, which cover the extremes of the vowel space. The aim is to find a spectral distance measure that best predicts when discontinuities are audible in order to find out how the diphone database can best be extended with context-sensitive diphones. To this end, we conducted a listening experiment and correlated the results with several measures of spectral distance.

*The work by Klabbers is part of the Priority Programma Language and Speech Technology (TST), sponsored by NWO (The Netherlands Organization for Scientific Research).

2. LISTENING EXPERIMENT

Five subjects with backgrounds in psycho-acoustics or phonetics participated in the listening experiment. The material was composed of 2284 C_iVC_j stimuli, which were constructed by concatenating diphones C_iV and VC_j excised from nonsense words of the form *C@CVC@*[1]. The stimuli consisted of five vowel conditions /a/, /A/, /i/, /I/ and /u/ in the context of all consonant pairs that can occur in C_i and C_j position. Preliminary tests showed that discontinuities and other effects in the surrounding consonants would overshadow the effects in the vowel. Therefore, the surrounding consonants were removed. In addition, the duration of the vowels was normalized to 200 ms and the signal power of the second diphone was scaled to equalize the level of both diphones at the boundary. The stimuli were randomized and the subjects were instructed to ignore the vowel quality and focus on the diphone transition. Their task was to make a binary decision about whether the transition was smooth (0) or discontinuous (1). The experiment was divided into six blocks, presented in three hourly sessions with a short break between two blocks. The block order was different for all subjects. Table 1 shows the percentage of discontinuities perceived in each vowel. A transition was marked as discontinuous when the majority of the subjects (80% or 4 out of 5) perceived it as such.

Vowel	Majority scores
/a/	17.1%
/i/	43.1%
/A/	52.1%
/I/	55.5%
/u/	73.9%

Table 1: Percentages of perceived discontinuities

The results show that the number of audible discontinuities in the /a/ is much lower than in the other vowels, whereas in the /u/ it is much higher. This is in line with findings in [10], where it was found that the largest amount of spectral coarticulation occurred in the /u/. Our results also reveal a slightly better score for the long vowels /a/ and /i/ than for the short vowels /A/, /I/ and /u/.

3. ANALYSIS

The results from the listening experiment were correlated with six spectral distance measures taken from various fields of research. They were used to measure distances across diphone boundaries between spectral envelopes. The Euclidean distance between (F_1, F_2) pairs, or Formant Euclidean Distance (FED), is often used in phonetics, the Kullback-Leibler measure (KL) [3] comes from statistics. The Euclidean distance between Mel-Frequency Cepstral Coefficients (MFCC), the Likelihood Ratio (LR) and the Mean-Squared Log-Spectral Distance (MS LSD) are used in automatic speech recognition [9]. The Loudness Difference (LD) and the Excitation Difference (ED) [6] come from the area of sound perception. All measures, except the formant Euclidean distance, were calculated from LPC coefficients computed over a 40-ms hanning window. The formant Euclidean distance and the Kullback-Leibler measure, will be discussed in more detail below.

In phonetics it is quite common to describe coarticulation in terms of the formants F_1 and F_2. In this investigation, the formants were measured by hand at the diphone boundary, which in our database lies 40 ms into the vowel for short vowels and 80 ms into the vowel for long vowels, which is typically not in the middle of the vowel. Close inspection of the stimuli reveals that most formant trajectories are fairly stationary throughout the vowel, except when the surrounding consonants are the alveolars /j/, /nj/, /tj/, /S/ and /Z/. Figure 2 displays the F_1 and F_2 values for the five vowels measured in our diphone database at the indicated locations. It shows that the /a/, /i/ and /I/ have small variations, whereas the /A/ and /u/ seem to be affected to a greater extent by their surrounding consonants. Especially for the /u/ differences in F_2 are considerable. They can be as large as 1200 Hz. The formant Euclidean distance is calculated by

$$D(i,j) = \sqrt{(F_{1,i} - F_{1,j})^2 + (F_{2,i} - F_{2,j})^2}$$

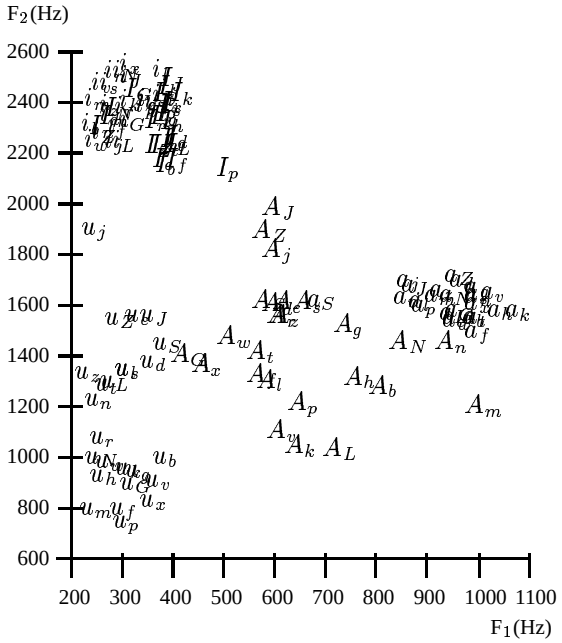

Figure 2: Vowel-subspace for /a/, /A/, /i/, /I/ and /u/

The Kullback-Leibler distance is a measure taken from statistics, which is used to compute the distance between two probability distributions $f(x)$ and $g(x)$. It is given by

$$\mathrm{KL}(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x.$$

In this study it is used to compare two power-normalized spectral envelopes. It has the important property that it measures logarithmic differences and that it weighs the difference with the power of the signal.

Our approach is that the subjects decide whether or not there is an audible discontinuity (by a 4 out of 5 decision). A spectral distance measure is said to predict a discontinuity when its outcome

[1] i and j are indices indicating the left consonant context and the right consonant context, respectively.

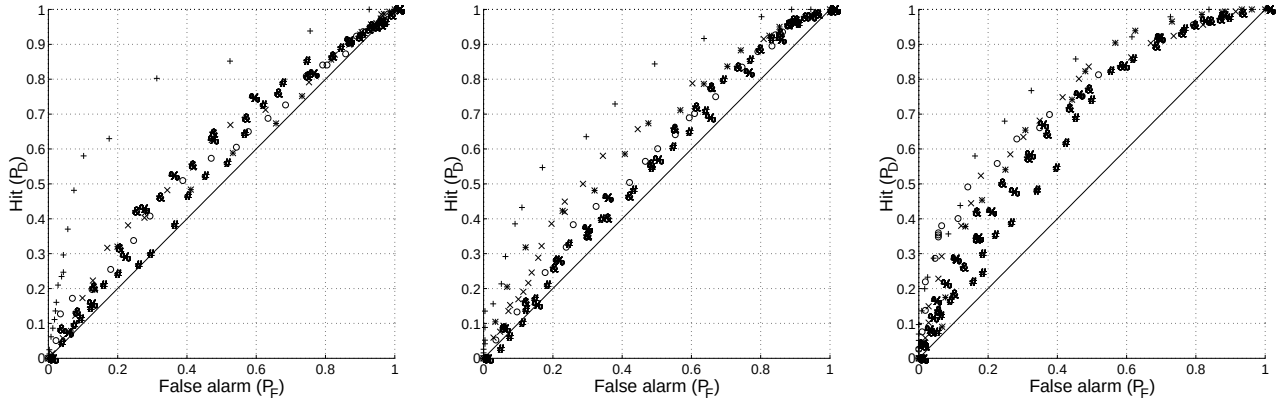

Figure 3: ROC curves for /a/ (left), /i/ (middle), and /u/ (right). *FED o, KL +, MFCC #, LR %, MS LSD &, LD x, ED **

D exceeds a predefined threshold β. The first thing that we want to investigate is to what extent the spectral distance measures are in agreement with the subjects' decision. Our second interest is to use the outcome of the best spectral distance measure to determine which additional context-sensitive units have to be added to the data base. This is related to the choice of β. We will postpone the discussion on β and first look at the usability of the spectral distance measures.

The measures and scores are correlated as follows. Firstly, the probability density functions $p(D|0)$ and $p(D|1)$ of the spectral distance D given that the transition was marked smooth or discontinuous, respectively, are estimated from the data. For a certain threshold β, the probability of a false alarm, the case that a transition is wrongly classified as discontinuous, is $P_{\mathrm{F}}(\beta)$ and the probability of detection (of a discontinuity) is $P_{\mathrm{D}}(\beta)$, which are given by

$$P_{\mathrm{F}}(\beta) = \int_{\beta}^{\infty} p(D|0)\mathrm{d}D, \quad P_{\mathrm{D}}(\beta) = \int_{\beta}^{\infty} p(D|1)\mathrm{d}D.$$

A plot of pairs $(P_{\mathrm{F}}(\beta), P_{\mathrm{D}}(\beta))$ for all values of β constitutes a Receiver Operating Characteristic (ROC, [5]). ROC curves are upward concave. Generally, a detector based on a given spectral distance measure is better than one based on another spectral distance measure, when at a fixed P_{F} it has a higher P_{D} or when at a fixed P_{D} it has a lower P_{F}. This means that the best ROC curve lies above and to the left of the others.

Figure 3 displays three ROC curves for the /a/, /i/ and /u/, whose inspection leads to a number of interesting observations. What can be seen is that the performance rating of some measures is not consistent across all vowel conditions. The formant Euclidean distance (o) performs well for the /u/, but worse for the other vowels. We previously saw that the formants in the /u/ are more affected by the spectral characteristics of the surrounding segments than the other vowels. Given that the formant Euclidean distance is a poor predictor for the other vowels implies that other factors besides formant differences are relevant[2].The Kullback-Leibler distance (+) qualifies as the best measure in all vowel conditions. The Euclidean distance for MFCC (#) performs very badly in all conditions, just above chance level. This is a logical finding, since it is a measure used frequently in automatic speech recognition, where it is meant to be insensitive to variations in different tokens of the same phonemes. The other measures perform quite well for the /u/ but moderately for the other vowels. Since no measure predicts the subjects' decisions perfectly, it seems plausible that other, possibly non-spectral, factors play a role in this phenomenon. Additionally, although the variability in the subjects' responses has been reduced by employing majority scores, there may still be some variability left.

[2]Incorporating F_3 and F_4 in the Euclidean distance calculation did not improve the prediction accuracy of this measure.

Now we come back to the postponed discussion on the choice of the threshold β. Our solution to the problem of discontinuities is to extend the diphone inventory with context-sensitive diphones in such a way that audible discontinuities are unlikely to occur. One way of doing this is by clustering.

Suppose we divide the diphone sets C_iV, $i = 1, \ldots, M$, and VC_j, $j = 1, \ldots, M$, for a particular vowel V into two sets of N clusters $\{L(V)_1, \ldots, L(V)_N\}$ and $\{R(V)_1, \ldots, R(V)_N\}$, such that the average spectral distance across diphone boundaries in corresponding clusters $L(V)_k$ and $R(V)_k$, $k = 1, \ldots, N$ is below a threshold β. Such a division can for example be done with a variant of the LBG algorithm [4]. The average distance between non-corresponding clusters $L(V)_k$ and $R(V)_l$, $k \neq l$ will then be greater than β. We now construct additional clusters $R(V)_{l,k}$, $k \neq l$, which contain the diphones of $R(V)_l$, but recorded with a left-side context consisting of a representative diphone in $L(V)_k$, e.g. the diphone closest to the centroid of $L(V)_k$. Instead of concatenating a diphone from $L(V)_k$ with one from $R(V)_l$ a diphone from $R(V)_{l,k}$ will be used, which will reduce the average spectral distance across diphone boundaries to approximately β. This procedure will increase the inventory size by a factor N.

If M is large enough, the fraction of spectral distances below β is given by

$$\int_0^{\beta} p(D)\mathrm{d}D = \int_0^{\beta} (p(D|0)P\{0\} + p(D|1)P\{1\})\mathrm{d}D,$$

in which $P\{0\}$ and $P\{1\}$ are the probabilities with which a transition is marked as smooth or discontinuous, respectively. These

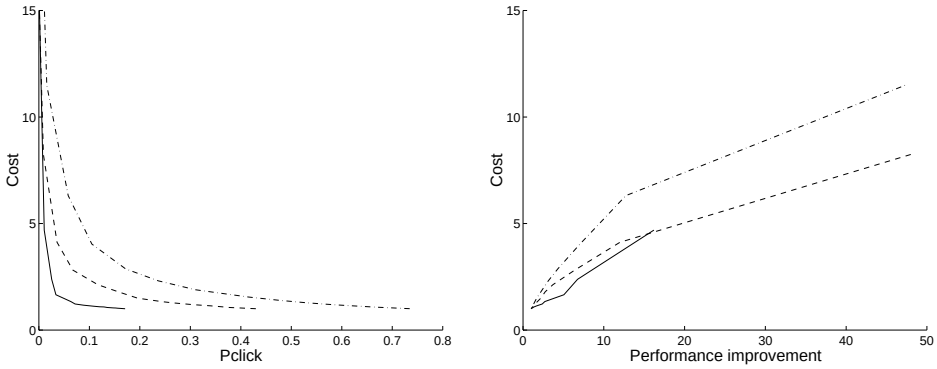

Figure 4: Performance and cost of /a/ (–), /i/ (- -), and /u/ (-.-)

probabilities can be estimated from the data. This implies that the number of clusters N, and therefore the factor by which the VC_j set needs to be increased, is given by

$$N(\beta) \simeq \frac{1}{\int_0^\beta (p(D|0)P\{0\} + p(D|1)P\{1\})\mathrm{d}D}.$$

This factor can be used as a measure of cost of improvement. After the extension of the diphone inventory, the probability of an audible discontinuity is given by

$$P_{\mathrm{click}}(\beta) = (1 - P_{\mathrm{D}}(\beta))P\{1\} = \int_0^\beta p(D|1)P\{1\}\mathrm{d}D.$$

This will be adopted as a measure of performance. Performance improvement can be expressed as $P\{1\}/P_{\mathrm{click}}(\beta)$. Performance (improvement) and cost can be related to one another by plotting pairs $(P_{\mathrm{click}}(\beta), N(\beta))$ or $(P\{1\}/P_{\mathrm{click}}(\beta), N(\beta))$ for various values of β (see Figure 4). The threshold β can now be chosen according to cost or performance constraints. What these figures show for example, is that the probability of an audible click for the /a/ in the original diphone inventory is the lowest, approximately 0.17, and in order for the /u/ to reach the same probability, the number of /u/C_j diphones has to be tripled.

4. RESULTS AND CONCLUSION

Our findings can be utilized to improve the quality of diphone synthesis. It supports suggestions to extend the diphone inventory with a number of context-dependent diphones in such a way that the occurrence of large spectral distances across diphone boundaries is avoided. For instance, the aforementioned problem with the word /zuk/ will be reduced if, in comparison to a word such as /kuk/, a different /uk/ diphone is used that better matches the preceding context. Diphone clustering can limit the number of additional diphones. Even though the Kullback-Leibler distance does not predict with 100% accuracy, it is the best measure at hand, and it is good enough to be used for clustering.

The experiments were confined to five vowels. In future also consonants will have to be investigated. In [7] many insights are provided into where spectral discontinuities are likely to occur. The liquids /l/, /j/ and /w/ are shown to be very susceptible to the spectral characteristics of the surrounding phonemes. Fricatives may show differences in high energy regions, due to the context. In diphone synthesis especially problems will occur when a voiced fricative is less voiced and more noisy in one context than in the other.

5. REFERENCES

1. A. Conkie and S. Isard. ”Optimal coupling of diphones”. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 293–304. Springer-Verlag, New York, 1997.
2. T. Dutoit. *An introduction to text-to-speech synthesis*. Kluwer Academic Press, Dordrecht, 1997.
3. S. Kullback and R. Leibler. ”On Information and Sufficiency”. *Annals of Mathematical Statistics*, 22:79–86, 1951.
4. Y. Linde, A. Buzo, and R. Gray. ”An Algorithm for Vector Quantizer Design”. *IEEE Transactions on Communications*, 28(1):84–95, 1980.
5. R. Luce and C. Krumhansl. Measurement, scaling and psychophysics. In S. Stevens, editor, *Handbook of experimental psychology*, chapter 1, pages 3–73. Wiley, New York, 1988.
6. B. Moore, B. Glasberg, and T. Bear. ”A Model for the Prediction of Thresholds, Loudness and Partial Loudness”. *Journal of the Audio Engineering Society*, 45(4):224–239, 1997.
7. J. Olive, A. Greenwood, and J. Coleman. *Acoustics of American English Speech: A dynamic approach*. Springer Verlag, New York, 1993.
8. J. Olive, J. van Santen, B. Möbius, and C. Shih. ”Synthesis”. In R. Sproat, editor, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, pages 192–228. Kluwer Academic Publishers, Boston, 1998.
9. L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.
10. H. van den Heuvel, B. Cranen, and T. Rietveld. ”Speaker variability in the coarticulation of /a,i,u/”. *Speech Communication*, 18:113–130, 1996.