FROM DIPHONES TO ALLOPHONES:
FROM DATA TO RULES

L.F.M. ten Bosch1, R. Collier1, L. Boves2

1 Institute for Perception Research, P.O.Box 513, 5600 MB Eindhoven, the Netherlands
2 Department Language and Speech, Nijmegen University, the Netherlands

Abstract

A research project is presented in which we aim to design a speech synthesis model based on both the diphone and the allophone concepts, i.e. the data-driven and rule-driven approach for speech synthesis, respectively. At present, diphone concatenation for Dutch leads to more intelligible speech than when a rule-based allophone synthesis is applied, although the latter synthesis has the theoretical advantage of optimal parametric freedom. However, in rule-based synthesis the iterative adjustment of the relevant speech parameters appears to be rather laborious. Therefore, an attempt is made to extract information about rules for allophone synthesis from the data-driven diphone synthesis (data-to-rule conversion) by means of a semi-automatic algorithm.

Aim of the project

The development of allophone synthesis of speech is a laborious task. The adjustment of perceptually relevant parameters takes much effort and often requires tedious comparisons between natural and synthetic utterances. Rules governing the perceptual quality of speech are often hard to find. In the allophone synthesis of Dutch, rules are developed by adjusting parameters and evaluating the resulting speech output in an iterative process (1). Presently, reasonable intelligibility has been obtained for all phonemes in nonsense CVC and VCV contexts, with a slightly poorer performance in the case of /k/, /d/, /x/ and a few dental fricatives.

On the other hand, the Dutch diphone synthesis (2) results in more natural and more intelligible speech, although the recognition of plosives and nasals is still slightly unstable in nonsense contexts (3,4). Roughly speaking, the quality of the diphone synthesis for Dutch lies between the quality of allophone synthesis and that of LPC-resynthesis.

From a linguistic point of view, allophone synthesis is much to be preferred over diphone concatenation. Phonological rules with respect to assimilation, insertion and deletion can easily be dealt with in rule-based speech synthesis, whereas diphones are, in principle, not subject to contextual alternation. Because of the better performance of diphone speech, it might be advantageous to look for methods to combine both synthesis principles. Diphones contain much information about the speech signal within very narrow contexts. Therefore, we propose that one look for methods to extract the relevant information from diphones in order to arrive at linguistically motivated allophone rules. If we can construct such methods, we can avoid tedious comparisons between natural and synthetic speech.

Stylization of diphone parameters

Diphone information is contained in speech parameter tracks. In the Dutch diphone synthesis system, the energy, voicing, and five formants and their bandwidths are coded and updated every 10 ms. These parameters represent 'raw data', and much of the temporal variation in a diphone is redundant. Therefore, diphones can be stylized to a certain extent without affecting the perceptual quality. In the present project, the diphones are stylized in such a way that they are made to correspond to a concatenation of allophone - transient - allophone as would be obtained in allophone synthesis. By doing so, we should be able to extract information about spectral targets and temporal organisation within diphones, and to apply the results to improve the rule-based allophone synthesis.

The stylization is performed by an iterative algorithm in a parameter space spanned by formants and bandwidths. After this stylization, a new diphone set is obtained in which each diphone is reduced to 2 targets (for 10 formant and bandwidth parameters, one gain parameter, and one voiced-unvoiced bit each) and three time markers \( t_1, t_2 \) and \( t_3 \). A stylized diphone defines for each parameter a constant target vector from \( t_0 = 0 \) up to \( t_1 \), a linear transition from \( t_1 \) up to \( t_2 \), and another constant target vector from \( t_2 \) up to \( t_3 \).

In order to perform such stylizations, we have to define:
The results mentioned here have been obtained while the parameter space was spanned by \( F_1, \ldots, F_s \) and the bandwidths \( B_1, \ldots, B_s \). The error criterion was chosen as the \( L_2 \)-norm with a weighting factor \( \frac{1}{\Delta t} \) for \( F_k \), \( k = 1, \ldots, 5 \), and a weighting factor \( \frac{1}{\Delta t} \cdot \Delta t \) for \( B_k \). The type of transient function was constant - linear - constant. The weighting \( \frac{1}{\Delta t} \) for \( F_k \) is inspired by the relevance of the bark-scale in auditory perception. As the bark-scale is logarithm-like from 500 Hz up to 2.5 kHz, absolute frequency differences \( \Delta(F) \) should be measured relative to \( F \) itself. In speech, the 'neutral' spectrum up to 5 kHz is characterized by the five formants \( \{F_1, 3F_1, 5F_1, 7F_1, 9F_1\} \).

The diphones that are used in this study contain 5 formant and 5 bandwidth parameters for the filter description. These parameters are obtained by a robust version of the Split-Levinson method for formant extraction (5). This robust formant extraction guarantees continuous trajectories in the parameter space.

Due to data reduction, the perceptual quality of stylized diphone speech will in general be less than that of original diphone speech. One cannot in a straightforward manner relate the loss of speech quality to the reduction of speech data: in order to do so, one needs knowledge of the relative importance of spectral and temporal cues in the speech signal. Some trends can be observed, however.

The sampling of speech parameters every 10 ms can be considered as a sampling process of a curve (trajectory) which is defined in the parameter space and determined by the utterance. Each updated speech frame (containing 10 LPC filter parameters) is then to be identified with a point in a 10-dimensional parameter space. An utterance (e.g. a diphone) can be considered as a trajectory in this parameter space, sampled at a 10 ms time interval. The process of diphone stylization is in fact equivalent to locally linearizing the involved speech trajectory.

For comparison, a second stylization has been performed in a space spanned by log area-parameters, which permit a satisfactory linearization of the speech curve by means of a temporal decomposition technique (6). Temporal decomposition (7) can be considered an alternative technique for finding spectral targets in such a way that the involved speech trajectory can locally be described as a linear combination of constant 'anchor' points. As a result, in the simplest case diphones are parametrized by two piece-wise linear interpolating functions and two target vectors; intermediate target vectors may also occur. The number of such targets that are necessary to give an adequate local description of the speech signal is given by the rank of a matrix \( M \). \( M \) is a matrix of speech parameters which appears in a natural way in the theory of temporal decomposition. Its rank can be evaluated by singular value decomposition (8). It appears that there exists a relation between the rank of matrices constructed from diphone data on the one hand and the perceptual quality of diphone speech as judged by listeners in informal tests: the smaller the rank, the poorer the speech quality (the reverse is not true). The stylization of diphones reduces the rank of the speech utterances to 1 (a diphone consisting of equal speech frames only would have rank 0).

Smoothling

After stylization, the spectral and temporal information contained in diphones is drastically reduced. The stylization does not automatically result in spectral matching at diphone boundaries, however. Therefore, a smoothing is applied to the whole set of diphones. This smoothing reduces the set of target vectors corresponding to the allophonic realizations of each phoneme to just one target vector, except for plosives, for which two targets are defined, corresponding with silence and burst.

From data to rules

After stylization and smoothing, a data-to-rule algorithm is applied. Presently, a (column) speech parameter vector \( a \), measured from the smoothed diphone set, is compared to phonological feature vectors corresponding to both phonemes in the diphone. The algorithm searches for phonological context dependence of the relevant speech coding parameters by looking for a weighting vector \( b \) such that

\[
\| a - Mb \| \quad (1)
\]

is minimized. In this formula, \( M \) denotes a matrix of phonological features (having -1, 0 and 1 as entries), and \( a \) a (column) speech parameter vector that has to be described in phonological terms. The result \( b \) indicates the weighting necessary for an appropriate interpretation in 'phonological' terms. Each row in the above matrix equation denotes a 'rule', describing the behaviour of the speech parameter vector \( a \) in terms of a phonological feature vector \( b \), while the feature matrix \( M \) is given.

Duration

It is known that additive duration rules in speech synthesis systems may not lead to optimal perceptual qual-
ity of phonemes in contexts, even when lower and upper bounds for durations are applied. Therefore, the durations are dealt with logarithmically before being subject to the above minimalization procedure (equation 1). The output of this procedure is dealt with exponentially.

Conclusion

It appears to be possible to synthesize intelligible speech by means of these very simple rules. Tested informally, the perceptual quality of speech made up of stylized diphones is somewhat less (on phoneme basis) than the quality of the original diphone speech. In particular, plosives and nasals perform poorly and should be improved. Yet, one of our conclusions is that much of the information in diphones can be reduced without seriously degrading the speech quality. In the presentation, we will show the results of formal listening tests on VCV and CVC nonsense words for both diphone sets.

Further research

Further research will focus on a reformatting procedure of these rules: several groups of phonological rules can be combined into only a few meta-rules by introducing free parameters in the rule format. Recently, the stylized and smoothed versions of diphones of two different speakers have become available. After having performed the above data-to-rule conversion we plan to look for the extraction of speaker-specific details.

Acknowledgement

Financial support of the Dutch national programme SPIN-ASSP is gratefully acknowledged.

References


