A NONLINEAR UNIT SELECTION STRATEGY FOR CONCATENATIVE SPEECH SYNTHESIS BASED ON SYLLABLE LEVEL FEATURES

Martin Holzapfel †*& Nick Campbell* ‡
†Siemens Corporate Technology.
ZT IK 5 81739 Munich, GERMANY
martin.holzapfel@mchp.siemens.de
‡ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

ABSTRACT

This paper describes an improved algorithm, motivated by fuzzy logic theory, for the selection of speech segments for concatenative synthesis from a huge database. Triphone HMM clustering is employed as an adaptive measure for articulatory similarity within a given database. Stress level contours are evaluated in the context of their surrounding vocalic peaks. The algorithm uses a beam search technique to optimise the suitability of each candidate unit to realise the desired target as well as continuity in concatenation.

1 INTRODUCTION

1.1 Motivation

Concatenative speech synthesis with a small fixed inventory suffers from two major short-comings: a) from signal degradation due to heavy post-processing (e.g., to enforce desired prosodic properties), and b) from discontinuities at the concatenation points. These problems can be significantly reduced by the selection of suitable units from a sufficiently large corpus [1]. The need to match a localised unit target, as well as the requirement to fit smoothly into a sequence of units can both be satisfied by the unit- and continuity-distance concept [2]. (see Figure 1)

The (unit and continuity) targets have to be matched in the both the prosodic and the phonemic domains. As the algorithm is currently implemented, search is performed on phone sized units of the database, and although the major acoustic events of the speech may be reflected adequately by the phonemic description, the finer details of articulatory variance cannot be modelled unless a wider context

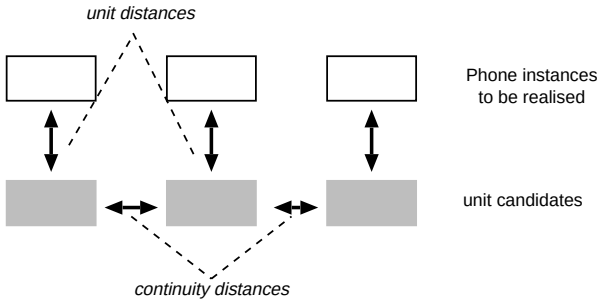

Figure 1: unit and continuity distance concept for unit selection

of at least the adjacent vocalic peaks is considered. For this reason, we are testing a system under which syllable-level attributes can be marked as features on the vowels.

1.2 Three Stage Selection Process

A run-time search for the optimal sequence of candidate unit segments can be pruned efficiently at the following three levels of computation:
1.) Preselection: When searching for the acoustical realisation of a given target speech segment, the units to be considered can be pre-selected according to their phonemic context. The prosodic target distance need be calculated only for units that are within the right classes of phonemic context. 2.) Filtering: Thresholds are calculated for a small number of prefiltered units so that only the subsequent instances likely to be realised need be considered as candidates for the (more expensive) calculation of the continuity at the concatenation points. 3) Path-determination:

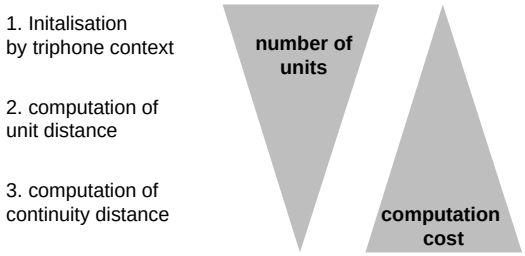

Figure 2: number of units versus computation cost in the three stage unit selection process

Only the surviving units, with their known target and concatenation distances, need be searched to find the global optimal unit sequence by a Viterbi Algorithm using a limited number of paths.

As illustrated in Figure 2, the number of units to be considered decreases significantly at each step, but the computing load is approximately constant throughout. Whereas the the computional cost for the distance calculation can be performed by a simple table lookup in Step 1, the processing load increases to a complex signal processing quadratic in the number of candidate units in Step 3.

2 ACOUSTIC CLUSTERING & MISSING DATA

In order to evaluate the similarity of phonemic contexts, a set of triphone HMMs is trained on the database [3]. After training separate HMMs for all the triphone contexts occurring in the database, these are clustered in a tree based manner. For this purpose all contexts of each phone are pooled first. This pool is subsequently split up according to phonetically motivated criteria, such as place of articulation, voicing of the context phones etc. A maximum likelihood criterion ensures maximum improvement in HMM modelling of the database with every split of a cluster [4]. The splitting proceedes until a threshold for increase in modelling improvement or the number of remaining units in the cluster is reached.

This minimum number of units in one cluster is balanced to compromise two contradictory requirements: on the one hand we need a broad prosodic and wide range contextual variety for selection in the later stages, and on the other, a tight spectral selection to avoid fuzziness in speech. For a 3 hour database containing 120k units a set of 3k clusters containing a minimum of 20 units each was found to be working properly.

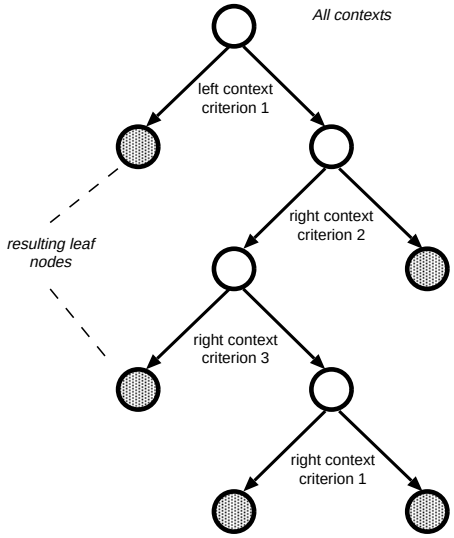

Figure 3: Split of HMM clusters according to phonetic criteria

This clustering does not only reflect the typical contextual variation and degradation of phones but is also adaptive to the manner of articulation of the database speaker. It allows the mapping of triphone contexts not occurring in the database by classification of their contexts according to the criteria learnt by the tree. The phonetic criteria that are spanning the clustering tree can by applied as a binary decision tree for any arbitrary context. This classification identifies the leaf node of the tree containing the units to represent the unseen triphone context.

Computing the extensive clustering operation offline reduces the check for the acoustic similarity of triphones to a simple table-lookup, which is performed at runtime in Step 1 of the selection proces.(see section 1.2)

3 SYLLABLE-BASED PROSODY FEATURES

Following Öhman [8], we view the main 'carrier' of speech as the vocalic stream ($v\prime$), interspersed with distinguishing consonantal information ($c\prime$). We can thus parametrically encode (or index) waveform characteristics by using a binary bit-vector having the granularity of the syllable. We envisage two tiers of phonation, each having effects on the other, but since the prosodic environment has stronger relations with the vocalic tier, it is the syllabic peaks ($v\prime$) that form the core of our index and carry information relating to the syllable as a whole.

For waveform concatenation we only need to locate appropriate 'centres' in each tier. Rather than joining at absolute or predetermined segment boundaries, we join between centres at the point of minimal discontinuity in the overlap of their transitions. Since in the best case the transitions out of the vocoid centre will be identical to the transitions into the contoid centre (and vice versa) these joins should be imperceptible.

The syllable peak ($v\prime$) is well described in low-dimensional space (see the IPA vowel triangle, or F1/F2 formant plots for example) but requires prosodic annotation for a better specification of acoustic vowel quality. Loudness, duration, F_0, and spectral-tilt are not features that need be labelled on the syllable per se, but can be predicted from the prosodic environment, which in turn can be largely determined from another bi-level system of peaks and troughs: the prominences and accents marking the focal structure of an utterance, and the phrase and clause boundaries delimiting its chunks.

The contoid tier ($c\prime$) is also well described by a small number of features like 'strength of intrusion' (weak: approximants, medium: fricatives, strong: plosives), and 'place of articulation' (front:labial, mid:palatal, back:velar), but subject also to influence from the vocoid tier and its prosodic modulations.

Prosodic information can be stored on this structure as a feature that need only be coded once per syllable. Since we encode the two interacting tiers as a sequence of syllable entries, marked on the vowels in the main index, it is only necessary to characterise each 'syllable' once by a $v\prime - c\prime$ pair. By assuming a 'silence' syllable at the beginning of each utterance, the onset characteristics of each subsequent syllable can be derived from the $c\prime$ of the previous. In this way, prosodic features spanning several neighbouring syllables can be easily accessed regardless of the number of intervening consonants (which can be as high as seven for a language like English).

4 FUZZY-LOGIC FOR THE SUITABILITY FUNCTION

4.1 Formal Framework

Evaluating the target distance for a candidate unit, or the distance between a pair of units to be concatenated, returns only a physical measure of the distance separating the two speech waveform signals, and is not necessarily a true indicator of the distortion that may be perceived when using the particular candidate units for speech synthesis.

The goal of the current approach is to try to find a function relating the physical distances measured from the signal to the suitability of the unit for synthesis in a given context, in a manner analogous to the perceptual stimuli relationship. This modelling concentrates all assumptions about the capabilities of the subsequent signal processing routines, and about human audio-perception etc., at this central point.

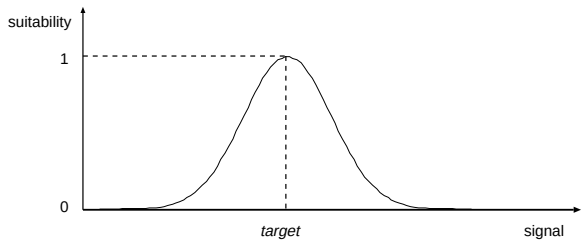

Figure 4: Fuzzy logic motivated suitability function for each partial target distance

The mathematical framework for the computation and combination of those suitabilities is motivated by fuzzy logic [7]. Assuming that the perfect sequence of units is not contained in a finite database, calculated mismatches between existing units have to be gradually quantised, balanced and a compromise found.

Using a procedure analogous to fuzzy membership functions, the suitability functions for each single distance are defined in the range of zero to one, where '1' denotes a perfected match, '0' an unacceptable mismatch, as illustrated in Figure 4. A typical suitability function e.g., for pitch target mismatch, could be a Gaussian over the pitch axis centred at the target pitch. Small (perceptually irrelevant) distances close to the target value result in suitabilities close to one. Big unacceptable mismatches result in zero suitability. Other units which are not perfectly matching but are within an acceptable distance are gradually weighted by the monotone decline between the extreme values. (Any similar monotonically declining function would match these assumption as well)

$$S_{global} = \prod_{units} \prod_{criteria} S_{partial} \qquad (1)$$

All isolated partial distances $S_{partial}$ are combined to one global suitability S_{global} for a possible sequence of units. All suitabilities within the set for each unit and along the path of units are multiplied (As denoted in Equation 1). Formally analog to the logical AND this combination demands each particular suitability to be in the accepted range, and emphasises the influence of big mismatches. While a linear com-

bination of the criteria calculates the mean suitability of a sequence of units, the multiplication ensures that one unacceptable unit will zero the suitability of the whole sequence. The prominence of a particular criterion is reflected by the shape of its suitability function. The bigger the range of units with high suitability for one criterion, the smaller the influence of this criterion on the final combined suitability.

4.2 Tuning of the suitability functions

There are two main approaches for the setup and the tuning of the suitability functions. The first one is driven by a-priori knowledge on the partial target distance (e.g., from perceptual experiments on the limitations of pitch manipulation by the PSOLA algorithm). A second approach is to assume a general shape for all employed partial suitabilities. Then the overall system output is optimised by relative shifts in the influence of single selection criteria. Each shift can be achieved by relatively narrowing or widening the shape of the related suitability function.

In our experiments we found a hybrid approach to be useful. The partial suitability functions are initialised by experiment (when available) or based on heuristically plausible figures. This set of initialisations is then balanced and adjusted to the database by optimising the system output according to subjective perception.

5 RESULTS

The algorithm described here was integrated and tested in two concatenative synthesis systems: CHATR (ATR) without and PAPAGENO (SIEMENS) with signal processing in the subsequent unit concatenation. Within PAPAGENO the proposed method showed a clear improvement, within CHATR informal listening tests using a forty-minute speech database [5] failed to show a significant preference in comparison with a highly optimised linear algorithm.

The conclusion that we draw from this is that the improved unit selection enables us to find candidate waveform segments that are closer to the desired acoustics, and thereby to reduce the amount of signal processing required. This results in an improvement in the quality of the PAPAGENO synthesis. However, in the case of CHATR, where there is no subsequent signal processing, there still remain some discontinuities between the segments due to the limited size of the source database.

6 DISCUSSION & FUTURE WORK

The quality of the above unit selection algorithm critically depends on the set of partial suitability functions and their interdependencies. Future work will aim to elaborate automatic training strategies for the suitability functions, as the manual tuning of these is highly complex and can be unstable to a single mistuned function.

The realised HMM clustering returns a binary decision on the similarity of two triphone contexts. This is appropriate to initialise the search algorithm but when computing the local unit suitability a finer distinction is desirable. This finer measure could be generated by further clustering in separated sub trees within each cluster. The distance within the nodes of each sub tree could then be used to evaluate acoustical distances within one cluster.

Acknowledgements

Thanks to Sabine Delinge, Yoshinori Sagisaka and Dave Sann for stimulating discussions.

References

[1] Hauptmann, A. G.(1993) "A First Experiment in Concatenation Synthesis from a large Corpus", Proc. Eurospeech, Berlin.

[2] Black, Alan W and Campbell Nick,(1995) "Optimising the selection of units from databases for concatenative synthesis", Proc. Eurospeech, Madrid.

[3] Donovan, R. E. (1995) "Automatic Speech Synthesiser Parameter Estimation Using HMMs", Proc. ICASSP, Detroit.

[4] Young, S. J.,Woodland, P. C., and Byrne, W. J. (1994) "Tree-Based State Tying for High Accurancy Acoustic Modelling", Proc. ARPA Workshop on Human Language Technology , Plainsboro.

[5] http://www.itl.atr.co.jp/chatr.

[6] Fant, G. (1991) "What can basic research contribute to speech synthesis?", J. Phon. 19, 75-90.

[7] Zadeh, L. A.(1965) "Fuzzy sets", in Information and Controll Nr. 8.

[8] S. Öhman "Coarticulation in VCV utterances: spectrographic measurements", Journal of the Acoustical Society of America 39, 151-168. 1965.