OPTIMISING SELECTION OF UNITS FROM SPEECH DATABASES FOR CONCATENATIVE SYNTHESIS

Alan W Black *Nick Campbell*

ATR Interpreting Telecommunications Research Laboratories.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
awb@itl.atr.co.jp nick@itl.atr.co.jp

ABSTRACT

Concatenating units of natural speech is one method of speech synthesis[1]. Most such systems use an inventory of fixed length units, typically diphones or triphones with one instance of each type. An alternative is to use more varied, non-uniform units extracted from large speech databases containing multiple instances of each. The greater variability in such natural speech segments allows closer modeling of naturalness and differences in speaking styles, and eliminates the need for specially-recorded, single-use databases. However, with the greater variability comes the problem of how to select between the many instances of units in the database. This paper addresses that issue and presents a general method for unit selection.

1. INTRODUCTION

The ATR ν-talk system for Japanese [4] efficiently selects non-uniform units from a large speech database but is specific to Japanese. English has more phonemes and more varying prosody, so a simple translation of the ν-talk system to English was not successful. A more general system is described here which deals with both English and Japanese and is designed to apply to other languages too.

This paper is concerned with only a small part of the whole speech synthesis process. Text processing (parsing, tagging, phrasing etc.) and linguistic processing (prosody and segmental prediction) are not discussed here. The third stage in this model, waveform synthesis, consists of two parts, unit selection and signal processing. Although this paper discusses how to select the best possible units, because the database used for selection will always be finite, even the best selection will not in general exactly match the desired utterance. Further signal processing will be required to modify the selected units. We are currently using PSOLA-based techniques [5] to modify the final selection though that aspect is not discussed here. However the closer the selected units are to the target segments the less signal processing will be required—and hence less distortion will be introduced.

[1]The term "concatenative synthesis" in used in this paper to mean concatenation of typically *sub-word* units of natural speech and not concatenation of words and phrases often used for synthesis when only a limited number of utterances are required.

2. UNIT SELECTION MODEL

In this system, each instance of a unit in the database (typically a phone-sized segment) is labelled with a vector of features. Features may be discrete or continuous. Typical features are phoneme label, duration, power, and F_0. Also, acoustic features such as spectral tilt are included in some of our databases. Other features describe the context of the unit: phoneme labels of neighbouring units, position in phrase, or direction of pitch/power change, etc. Note also that vectors may include features about a unit's context as well as the unit itself. Where possible, features are described in a normalised form, e.g. distance in standard deviation units around a zero mean [2]. A further requirement is that there be a distance measure between two feature values of the same type. For continuous features this is easier, but for discrete features (e.g. phonemes) a distance needs to be explicitly defined. Distances between feature values are normalised to lie in the range 0 (good) to 1 (bad).

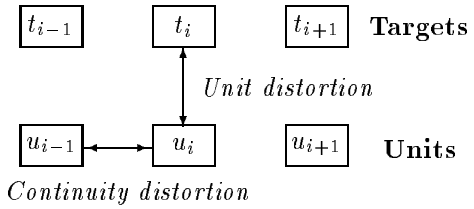
For selection, the target segments, predicted by earlier components of the synthesizer, or for testing purposes taken from natural speech, are also labelled with a subset of these features—specifically excluding any features only available from acoustic measures. These explicitly specify the intended utterance's segmental and prosodic characteristics.

In order to measure how well a set of selected units match a set of target segments, two types of **distortion** can be defined.

Unit distortion $Du(u_i, t_i)$ is defined as the distance between a selected unit and a target segment, *i.e.* the difference between the selected unit feature vector $\{uf_1, uf_2, ..., uf_n\}$ and the target segment vector $\{tf_1, tf_2, ..., tf_n\}$ multiplied by a weights vector W_u $\{w_1, w_2, ..., w_n\}$.

Continuity distortion $Dc(u_i, u_{i-1})$ is the distance between a selected unit and its immediately

adjoining previous selected unit, defined as the difference between a selected unit's feature vector and its previous one multiplied by a weight vector W_c. This distance represents the cost of joining two units. This vector includes the unit distortions of a selected unit's context with the unit context of the previous selected unit it is to be concatenated to.



Varying values in W_u and W_c allows the relative importance of features to change, for example to allow F_0 to play a greater role in unit selection than duration. The values may also be zero thus eliminating a feature from the selection criteria. The weights vectors W_u and W_c will be different.

The **best unit sequence** is defined as the path of units from the database which minimizes

$$\sum_{i=1}^{n}(Dc(u_i, u_{i-1}) * WC + Du(u_i, t_i) * WU)$$

Where n is the number of segments in the target utterance, and WC and individual WU are weights. Maximizing WC with respect to WU, minimizes the distortion between selected units at the expense of distance from the target segments.

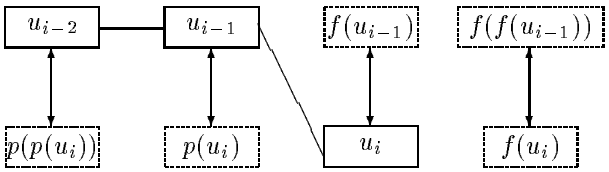
3. FEATURES USED

Although various additional features may be used, all our databases include the following features.

In unit distortion we use: phonetic context, duration, log power and mean F_0.

In continuity distortion three features are used: phonetic context, prosodic context (duration, power and F_0 together), and acoustic join cost. These are described in detail below.

In the case of prosodic and phonetic context, distances are taken between a window of four units around a join. Thus in joining u_i to u_{i-1}, prosodic and phonetic distances are taken for the vertical pairs.

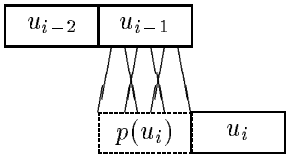


Non-dashed boxes represent selected units (connected by lines). Dashed boxes represented units not actually selected, but consecutive in the database with selected units. The function p returns a unit's actual previous unit *from the database* while the function f returns its actual following unit.

So far, a window of two preceding and two following units is used to calculate the continuity distortion. Unit distortions closer to the join are weighted more than the two further away.

Note that it is important that (at least) prosodic context is checked with respect to the preceding *selected unit* not just the target. For example a unit u_i whose F_0 is close (slightly lower) to target t_i may be chosen while the following selected unit u_{i+1} may be close to target t_{i+1} but slightly higher, thus introducing a larger distortion between u_i and u_{i+1} than may be necessary with a different choice of candidate.

A third distance used in measuring continuity distortion is an acoustic one. Quantisation (128) of mel-cepstrum (16 parameters) at 10ms frames is calculated for the whole database[2]. For each pair of units to be joined, an overlapping sliding window of seven frames (biased significantly into the previous selected unit) is searched for the closest "VQ" distance.



This gives both a score and an offset to the "best" possible join point for a pair of candidate units. Thus actual joins need not occur at the labelled boundary in a database, and typically occur towards the middle of the unit. Given a reasonably sized database $p(u_i)$ will be of the same phonetic type and prosodically similar to u_{i-1} where possible.

Although we have currently been testing with only the above features we do not see this as the complete set. Other features may easily be added to the distortion measures and as we refine the system we will add more. We also wish to investigate the significance of different features in order to minimise the amount of computation required.

4. SEARCH ALGORITHM

A Verterbi search (with beam-width constraints) is used to find the path with the minimum cost as defined by the expression above. An exhaustive search would be too computationally expensive.

Given a set of targets representing the utterance we wish to synthesize, for each target segment we find all units in the database with low phonetic distance from the target[3]. Next we find the unit distortion for these units with respect to the target, prune

[2] Other quantizations of different encodings may be performed but have not yet been investigated.

[3] Typically this means at least diphone matching though this initial selection guarantees there will be at least some candidates.

this list taking the m best ones (m is typically 20-50). Next for all these candidates find the continuity distortion between all candidates from the previous target. Prune this list to the n best costed pairs (n is typically 20-50—but need not necessarily equal m). Continue through all targets finding n paths at each stage. Select the best path at the end.

5. OPTIMISING THE WEIGHTINGS

The quality of selected units depends heavily on the weights for the various feature distances in both unit and continuity distortion. Although these weights can be tuned by hand, a more systematic method of tuning these produces better results.

Defining the optimal value of the weights so that the *best* selection produces the perceptually best quality synthesis is non-trivial. An objective measure is required to determine if the *best* selection is perceptually better to a hearer of the utterance.

However testing human perceptions is not easy, such tests are prone to errors. They are not very discriminatory at fine detail, and not suitable for large numbers of utterances (i.e. humans cannot reliably check thousands of examples). Instead we use the mean Euclidean cepstral distance [3, pp 150-171] between (time aligned) vectors from the selected units and the target segments (which are completely known in the special test case of mimicking natural speech utterances from the speaker of the source database), but again it is open to question how sensitive such a cepstral distance is, and how closely it correlates to human perception.

The following is the current method for optimising the weights, it is a first approximation and computationally intensive, but produces weights that produce better synthesis than hand tuned weights—in larger databases sometimes significantly so.

A natural utterance is selected from the database and is presented as a sequence of target segments, that utterance is removed from the database so none of its units are available for selection. For a large range of weights, the above beam search algorithm is used to find a best selection. Cepstrum frames for that best selection are time-aligned with the cepstrum frames from the original natural utterance. The mean Euclidean distance is calculated between the cepstral vectors of the selected units and the natural original of the targets. That distance is used as a score for the set of weights.

To avoid over-tuning for a particular utterance, a number of test sentences are used and their scored weights are compared. Weightings appearing high in many examples are considered good.

It should be stated that this method of optimising weights is very computationally expensive. On a high end workstation for testing 11 weights with 3 different values (around 20,000 tests) it takes about 24 hours. However once the "best" weights are found our synthesizer improves in quality and still runs in real time.

6. EVALUATION OF CEPSTRUM MEASURE

This cepstrum measure although not guaranteed optimal, does clearly work for major changes in weights. Selections with small cepstrum distances are much better than those with large cepstrum distances. However different selections with close cepstrum measures are frequently not humanly distinguishable.

In order to better evaluate the relationship between objective cepstral and subjective perceptual distance measures, we asked six subjects to score a set of differently weighted test utterances and correlated their averaged rankings against the acoustic measures. The test consisted of seven different weightings for two sentences presented three times in random order. The subjects were asked to count the number of "bad segments". In practice the magnitude of scores varied widely between subjects, but after counts were normalised, subjects were consistent with themselves and (mostly) agreed on the order of acceptability of the presented utterances, these results are discussed more fully in [2].

From this simple test we were able to make the following observation. The cepstral distance seems to give more importance to unit distortion at the expense of continuity distortion, while human perception favours more weight on WC (i.e. less continuity distortion). This is because the cepstral measure takes the mean error over each point. Therefore continuous closeness is favoured over short "burst errors" that occur at bad joins. Humans however are upset by burst errors as well as prosodic mismatches, and hence prefer a balance of WC to WU.

These tests were made before acoustic and phonetic context were added to continuity distortion measures. With these added, the number of "burst errors" do decrease but still the cepstral measure is less sensitive to "burst errors" than humans are. Although difficult to find a good automatic measure that directly follows human perception of speech quality, we are currently working to improve on our current simple cepstrum distance measure.

7. DATABASES

A synthesizer voice may be constructed fully automatically from waveform files and phone label files (although the process may take several hours). Some of our databases have phones labels automatically assigned from word labels using an aligner, thus making the process of database construction require less skilled work. Because we use acoustic measures and search for appropriate join points during selection, accurate phoneme boundaries are not very important.

We have already built a number of databases both in Japanese and English with which we have tested this selection system. [2] discusses the problems of labelling and pruning of databases but here we will only discuss a number of databases metrics. The databases currently created offer interesting dimensions in size and style. (Comments in parentheses indicate change with respect to the first database.)

sab200: 200 phonetically balanced sentences spoken by a female British English speaker—9,023 units.

sab5000 (different style) 5000 phonetically balanced isolated words spoken by the same British English female—37,716 units.

gsw200: (different gender) the same 200 phonetically balanced sentences spoken by a male British English speaker—9,012 units.

f2b: (larger) 116 utterances from an American English female news announcer, from BU-RADIO corpus—37,597 units.

mhtBset: (different language) 500 phonetically balanced sentences spoken by male (Tokyo) Japanese speaker—30,322 units.

After some informal listening tests of synthesis from these database we have made the following observations.

Because **sab5000** database contains much longer (mean durations are longer) and less intonationally rich raw units than **sab200**, the synthesis sounds slow and over articulated. **gsw200** produces better synthesis than the **sab200**, even though these databases are the same sentences. This suggests more consistency in this male speaker than the female speaker. The larger **f2b**-based speech synthesizer produces better speech as the smaller **sab200** database is often limited in alternative units. Finally the Japanese database **mhtBset** produces significantly better speech than any of the other databases. This is most probably due to it being male, a professional (very consistent) speaker, Japanese (where there are fewer phonemes and less prosodic variation between them), accurately labelled and a larger database.

8. DISCUSSION

Although our current model basically works there are still many areas that can be improved. The cepstrum measure as an objective measure of quality is a first attempt, but finding a optimal automatic measure of distance between waveforms is a non-trivial problem. However finding a better measure than a simple mean cepstrum distance should be possible.

Weight training is currently done by a simple full search of points in weights space. As we intend to increase the number of features, and allow variation in some currently hardwired parameters a better (faster) training algorithm is desired. There are possibilities and borrowing from work used in speech recognition looks promising.

All our tests have been on raw concatenation of selected units. Further signal processing (e.g PSOLA) is necessary in order to modify the selection closer to the desired targets, though the better the selection the less signal processing is necessary. Although signal processing may introduce different amounts of distortion depending on the type of modification (power, duration and F_0) we have not yet investigated this nor how our unit selection criteria might be modified to take advantage of the various costs in different aspects of modification.

In addition to improving the method, the number of features can be increased (and existing ones improved). Another dimension is in adding more databases (and pruning the size of existing ones). Although we have so far only tried two languages (English and Japanese) nothing in the method is language dependent, we have collected a Korean speech database and are currently labelling it and creating a synthesizer from it.

The unit selection algorithm described here is included in the CHATR speech synthesis system [1] and offers reasonable natural synthesis in real time.

9. CONCLUSION

A general model has been presented for unit selection from labelled natural-speech databases for synthesis by minimizing unit and continuity distortion. A method for tuning weights allowing varying importance of features is described which offers control over their optimal settings. It has currently been tested with both English and Japanese, for both male and female speech.

REFERENCES

[1] A. W. Black and P. Taylor. CHATR: a generic speech synthesis system. In *Proceedings of COLING-94*, volume II, pages 983–986, Kyoto, Japan, 1994.

[2] N. Campbell and A. Black. Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirshberg, editors, *Progress in speech synthesis*. Springer Verlag, 1995.

[3] L. Rabiner and B. Juang. *Fundamentals of speech recognition*. Pretice Hall, 1993.

[4] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR – ν-TALK speech synthesis system. In *Proceedings of ICSLP 92*, volume 1, pages 483–486, 1992.

[5] H. Valbret, E. Moulines, and J. Tubach. voice transformation using PSOLA technique. *Speech Communications*, 11:175–187, 1992.