Prosody and the selection of source units for concatenative synthesis

Nick Campbell
Alan W. Black

ABSTRACT
This chapter describes a procedure for processing a large speech corpus to provide a reduced set of units for concatenative synthesis. Crucial to this reduction is the optimal utilisation of prosodic labelling to reduce acoustic distortion in the resulting speech waveform. We present a method for selecting units for synthesis by optimising a weighting between continuity distortion and unit distortion. The source-unit set is determined statistically from a speech corpus by representing it as the set of sound sequences which occur with equal frequency, *i.e.*, by recursively grouping pairs of segment labels to grow non-uniform-length compound label-strings. Storing multiple units with different prosodic characteristics then ensures that the reduced database will be maximally representative of the natural variation in the original speech. The choice of an appropriate depth to which to prune the database reflects a trade-off between compact size and output voice quality; a larger database is more likely to contain a prosodically appropriate segment that will need less modification to reach a target setting in the concatenated utterance.

1 Introduction

As large corpora of natural speech are becoming more widely available, we can consider them not just as source materials for the modeling of language and speech characteristics, but also as a source of units for concatenative synthesis.

Concatenative synthesis systems have traditionally employed only a small number of source units, using one token, or waveform segment, per type of unit in the source-unit inventory. Such systems produce highly intelligible speech quickly and economically, while using only a small amount of computer memory and processing. However, as yet, they have failed to produce really natural-sounding speech.

Part of the reason that even concatenative synthesis still sounds artificial is that the source units for concatenation are typically excised from

recordings of carefully read lab speech, which although *phonemically representative*, is constrained to be *prosodically neutral*. The speech tokens used as source units thus encode the relevant static spectral characteristics (or configurations of the vocal tract) for a given sound sequence but fail to adequately model the different dynamic articulatory characteristics of that sequence when it is reproduced in different meaningful contexts.

The prosodic variations in human speech across different speaking styles or between different parts of the same utterance are normally accompanied by corresponding variation in phonation: *i.e.*, by changes in voice quality such as the emphasis of higher-frequency energy that comes from the longer closed phase of 'pressed voice' or the steeper roll-off of energy that comes with 'breathy voice'. Whereas current signal processing techniques used in concatenative synthesis are adequate to warp the prosody of source units to model coarse variations in fundamental frequency (f_0), duration, and energy, they fail to adapt the spectral characteristics of the concatenated units to encode these fine phonation differences and, as a result, the synthetic speech sounds artificial or hyperarticulated. The simple modification of prosody without an equivalent modelling of its phonation effects is insufficient.

Furthermore, since only a limited number of speech tokens are used in the generation of a large variety of speech utterances, considerable degradation can result from the amount of signal processing required to modify their prosody to suit the wide variety of target contexts. For example, in matching the duration of a unit, it is customary to repeat or delete waveform segments to artificially create a longer or shorter sound. This is inherently damaging to naturalness.

The solution we propose for the above problems requires a larger inventory of speech tokens for source units, allowing several tokens for each type[1] so that the token closest to the target context can be used in synthesis to minimise subsequent signal processing. Since designing such a large inventory can be difficult, and recording such a database very time-consuming, we have developed tools to process existing corpora of natural speech to extract suitable units for concatenative synthesis.

Extraction of a maximally varied and representative set of speech tokens from a given speech source requires three stages of processing: a) segmental and prosodic labelling of the speech corpus, b) analysis of frequencies and distributions of each segment type, and c) selection of a reduced-size but optimally representative set of source tokens to cover the variations encountered in each type.

---

[1] We will henceforth use the term *type* to refer to monophone, diphone, or polyphone *classes*, and the term *token*, to refer to actual *instances* of each type (*i.e.*, waveform segments taken from a speech database).

2 Segmental and Prosodic labelling

A basic requirement for a speech corpus to be processed is that it has an accompanying orthographic transcription. This can be aligned segmentally by generating the phone sequences that would be used to synthesise it and using Markov modelling to perform the segmentation. Manual intervention is still required in the segmental labelling.

Prosodic labelling is automatic, given the sequence of phone labels and the speech waveform. For each segment in the corpus, measures are currently taken of the following prosodic dimensions; these are derived automatically from combined output of the ESPS *get_f0* and *fft* programs [ERL93]. With the exception of 'duration' the measures are taken at 10msec intervals throughout the speech signal and then averaged over the duration of the phone-sized segments of speech as delimited by the labels:

- duration,
- fundamental frequency
- waveform envelope amplitude,
- spectral energy at the fundamental,
- harmonic ratio, and
- degree of spectral tilt.

These values are then z-score normalised for each phone class to express the difference of each segment from the mean for its type in terms of the observed variance for other tokens of that type for each of the above dimensions. To model the position of each token in relation to changes within the respective prosodic contours, first differences of these normalised values are then taken over a window of three phones to the left and right of each segment. The sign of the result indicates whether a segment is part of a rising contour (*e.g.*, increasing pitch, loudness or length) or falling. The magnitude indicates the rapidity of the change.

From this prosodic information (which is also later used to determine optimal units in the selection process for synthesis), we can discriminate between tokens in otherwise identical segmental contexts.

3 Defining units in the database

In a labelled speech corpus, the number of **types**, as defined by the monophone labels on the segments, is small (on the order of 20 to 50 for most languages) while the number of **tokens** of each type depends on the size of the corpus but also varies considerably between types, from very large (for

a few vowels) to extremely few (for some rare consonants). The type and token distributions are defined both by the language and by the contexts from which the speech data was collected; if the corpus is sufficiently large, then all sound types of the language will be represented, but the number of variants for some will be few.

The aim when processing such a corpus to produce synthesis units is to preserve all tokens of the rare types while at the same time eliminating duplicate or redundant tokens from the more common types. Since system storage space is often limited and, even with efficient indexing, searching for an optimal token in a large database can be very time-consuming, efficient pruning of the corpus is necessary in order to control the size of the source database, while maximising the variety of tokens retained from it.

As there are never likely to be two tokens with identical waveform characteristics, 'duplicate' is defined here to imply proximity in the segmental and prosodic space, as limited by the requirements of the storage available for units in the synthesis system. That is, subject to system capacity, we can store a number of different tokens of each type to ensure maximally diverse coverage of the acoustic space.

3.1 Segmental types and tokens

The relation between the prosody of an utterance and variation in its spectral characteristics has long been known [Gau89, Tra91]. Lindblom [Lin90] describes the continuum of hyper- and hypospeech observed in interactive dialogues, by which speakers tune their production to communicative and situational demands. Sluijter & van Heuven [4], also citing such work on overall "vocal effort" such as Gauffin & Sundberg [Gau89], showed that, in Dutch, stressed sounds are produced with greater local vocal effort and hence with differentially increased energy at frequencies well above the fundamental. More recently, Campbell & Beckman [Cam95] confirmed that for English too, spectral tilt is affected by linguistic prominence.

It is not yet easy to quantify such phonation-style-related differences in voice quality directly from a speech waveform, but fortunately the differences correlate with grosser prosodic features such as prominence and proximity to a prosodic-phrase boundary. To select a subset of tokens that optimally encodes the phonation-style variants of each segment type, we therefore adopt a functional approach and determine instead the contexts in which the prosody is markedly different. Because the weak effects of phonation co-occur with the stronger prosodic correlates, we label the strong to also encode the weak. Thus it is not necessary to be able to detect the changes in phonation style directly in a speech corpus, rather we can capture them from the gross and more easily detectable differences in f_0 and duration to encode the speech segments.

3.2 Determining the units

When reducing the size of a speech corpus to produce source units for concatenative synthesis, we need to store the most representative tokens of each type, ideally several of each, to be sure of enough units to cover prosodic variation, but no more than necessary to model the language. The first step is therefore to determine a set of types that uniformly describes the distribution characteristics of phones in the corpus.

Several methods have already been suggested for the automatic extraction of units for a speech synthesis database [Nak88, Sag92, Nak94], but these have concentrated on maximising the variety of segmental contexts, we emphasise here the importance of also considering the prosodic context.

Under the assumption that listeners may be more sensitive to small variation in common sound sequences (such as in function-words) and more tolerant of variant pronunciation in less common words or in names, we cluster the original monophone labels by frequency to determine the common sequences in the corpus. In this way, the number of 'types' is increased by clustering the original phone labels to form non-uniform-length sequences (compound phone-strings) for an optimal representation of the common sound sequences in the corpus.

As function words are very common in speech, they will tend to emerge more readily from the clustering process. As these words are often produced with special reduction in fluent speech, this clustering allows them to be automatically modeled separately as discrete units, without the need for special linguistic analysis of the corpus.

The algorithm for determining the unit set is given in pseudocode in Fig 1 is derived from [Sag92] but uses counts instead of entropy. In a process similar to Huffman coding, the most frequently-occurring label is conjoined with its most frequently co-occurring neighbour to produce a new compound type and the cycle is repeated using the increased set. At each iteration the number of types grows, and the token count of the two most frequent conjoined types correspondingly decreases.

The loop terminates when a threshold specifying the maximum number of tokens for any type has been reached. This threshold (and by implication, the ultimate number of types) is arbitrarily chosen, according to the initial size of the corpus and the degree of reduction required, to be approximately five-times the number of tokens required from each type. The greater the number of tokens per type, the better the prosodic flexibility of the final unit set.

This stage of processing yields a set of new labels which describe the frequency of co-occurrence of the speech sounds in the corpus. The resulting set of compound types ensures that (with the exception of the very few sparse-token types) each unit is equally likely in the language being modelled. Infrequently occurring types (such as /zh/ in English or /tsa/ in Japanese) will not be clustered, and all tokens of each are preserved.

```
main-loop
set threshold = n
initialize: types = labels

while( number_of_tokens( any type ) > threshold ) do
    current_type = max( count tokens of each type )
    for( most frequent type) do
        max = find_most_frequent_neighbour( current_type )
     return( compound_type ( current, max ))

find_most_frequent_neighbour( current )
    for( tokens in database )
        if( type == current )
        count left_neighbour types
        count right_neighbour types
    return( max( left_neighbour types, right_neighbour types )

compound_type ( current, most_freq )
    if( most_freq == left_neighbour)
        return( concat( most_freq_neighbour + current_label ))
    else
        return( concat( current_label + most_freq_neighbour ))
```

FIGURE 1. Rectangularisation algorithm: Subdivide the frequent units, and cluster them with their most common neighbours to form longer units.

The remaining tokens of the more common types can now be reduced to prune down the size of the database to meet system limitations. However, rather than simply select the one most typical token of each type, we preserve a number of each, up to the limits of storage space.

3.3 Pruning the database

To ensure best coverage of the prosodic space, we next select the n most prosodically diverse tokens to represent each type of unit. Vector quantisation is performed to cluster all tokens of each type in turn according to their prosodic characteristics. For each type, the tokens closest to the centroid of each cluster are then taken as representative.

The number of clusters (n) specifies the depth n to which to prune the database. Thus the ultimate size of the source-unit database will be slightly less than n times the number of types (there usually being less than n tokens of the rarest few types). The size of the quantization codebook is thus determined as a function of the number of tokens to be retained for each type.

The choice of an appropriate depth to which to prune a large source corpus is a trade-off between compact size and synthetic speech quality; a larger source-unit database is more likely to contain a prosodically appro-

priate segment that will need less modification to reach a target setting in the concatenated utterance. If only one token were to be stored for each type (*i.e.,* to produce the smallest and least flexible unit set that still covers the segmental variety in the corpus), then we would simply choose the token closest to the centroid for each type, without using prosodic vector quantisation, to be sure of having the most typical token.

By taking more than one token to represent each type, we are assured not only that all contextual segmental variation in the original corpus will be preserved. but also that we will be better able to match different prosodic environments; the more tokens, the more flexibility of coverage. In this way we no longer have to synthesise using the one most typical token in all prosodic contexts, but can select from several. The supposedly redundant units, whose segmental environment is the same but which actually differ in prosodic aspects, can be selected from to minimise the amount of waveform distortion needed to match the target prosody for a particular utterance.

4 Prosody-based unit selection

Each unit in the database is labelled with a set of features. These features include phonetic and prosodic features (such as duration, pitch power etc.) to which are added acoustic features (such as cepstral frame quantisation). The features available are database dependent, though we expect at least phone label, duration pitch and power. As far as possible features are specified in terms of z-scores so distances are normalized between features. Other features are used in a database unit description that do not directly affect selection (*e.g.,* position in the waveform file).

For selection of units for synthesis, the target segments (predicted by earlier components of the synthesizer, or for testing purposes taken from natural speech) are specified with a subset of these features to specify the characteristics of the utterance and its prosody.

Because of the richness of this information, we do not (though could if our databases were so labelled) use all the acoustic measures described in [Sag92] for selection. We are testing the assumption that appropriate prosodic characterisation will capture the acoustic differences in the units, and that we can thereby avoid computationally expensive acoustic measures to achieve faster unit selection.

In this selection model, we define two types of distortion to be minimized to find the best sequence of units.

- **Unit distortion** is defined as the distance $Du(s_i, t_i)$ between a selected unit and a target segment, *i.e.,* the weighted mean distance between non-zero-weighted features of the selected unit feature vector $\{sf_1, sf_2, ..., sf_n\}$ and the target segment vector $\{tf_1, tf_2, ..., tf_n\}$. Distances are normalised between 0 (good) and 1 (bad). Weights too

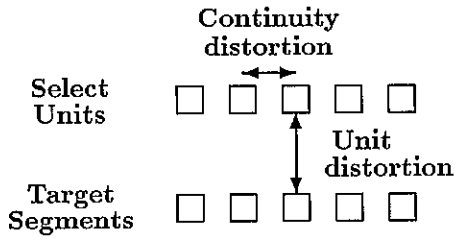
are between 0 and 1 causing Du to also lie between 0 and 1.

- **Continuity distortion** is the distance $Dc(s_i, s_{i-1})$ between a selected unit and its immediately adjoining previous selected unit. This is defined similarly to Du as a weighted mean distance between non-zero-weighted features of a selected unit's feature vector and that of the previous unit.

The weights for significant (i.e. non-zero-weighted) features in unit distortion and continuity distortion will be different, as will be the choice of features. Vectors may also include features about a unit's context as well as the unit itself.

Varying the weights allows the relative importance of features to change, for example allowing pitch to play a greater role in selection than duration. The values may also be zero, thus altogether eliminating a feature from the selection criteria. The weights for unit distortion will differ from the weights for continuity distortion.

The following diagram illustrates this distinction between distortion measures.



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

$$\sum_{i=1}^{n}(Dc(s_i, s_{i-1}) * WJ + Du(t_i, s_i) * WU)$$

where n is the number of segments in the target, and WJ and WU are further weights. Maximizing WJ with respect to WU minimizes the distortion between selected units at the (possible) expense of distance from the target segments.

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

Human perceptual tests are one measure but they are prone to errors and are neither very discriminatory at fine detail, nor automatic. Another more objective measure is the mean Euclidean cepstral distance [Rab93, pp 150-171] between (time-aligned) vectors of selected units and target segments. In the special test case of mimicking a natural speech utterance from the speaker of the source database, the target features are completely known

and an objective quantification of the degree of match can be obtained. But, it is open to question how closely the cepstral distance measure predicts human perceptual characteristics, This issue is addressed further below.

4.1 Optimising the weightings

A beam search algorithm was used to find the best units based on the above minimization criteria. Originally for each database all weights were hand tuned without any formal measurement (and interestingly they were noticeably different for different databases), The following procedure is now employed to optimise these weights automatically and determine the best combination for each database.

> First an utterance is removed from the database so its segments are no longer available for selection. Its natural segments are used to define the parameters of the test utterance to ensure that testing is independent of any higher levels of the synthesis system.
>
> The beam search algorithm is used to find the best selection of units that minimize the distance described above, with respect to a set of weights.
>
> The cepstra of the selected units are time aligned with those of the target segments, and the mean Euclidean cepstral distance between target (original) segments and the selected (replacement) units is calculated.
>
> The process is repeated with varying weights until they converge on a minimum mean cepstral distance.

This process is not ideal, as minimizing cepstral distance may not maximize the quality of the synthesized speech, but it is an objective measure which as we will show offers some correlation with human perception. Problems with this measure and some ways in how it may be improved are discussed below.

5 Evaluation

Unit selection is only a small part of the whole synthesis process, and for full synthesis, higher level modules generate a segment description specifying the appropriate features (*e.g.*, predicting values for f_0, duration, power), and subsequent signal processing is applied to the waveforms of the selected units modifying them to the desired target values and reducing any discontinuities at the joins between units.

However, for the tests presented here, we are concerned only with unit selection and did not want to confound our results with any errors from

higher-level modules. We therefore used as a target the prosody and phone sequences of original natural utterances, and performed no subsequent signal processing. In this way we were able to test the selection of units under controlled conditions without having to take into consideration possible effects of other modules.

5.1 Test 1: full database

Using a medium-size database of English female radio announcer speech (44,758 phone labels) [BU-95], we varied weightings of four individual features used in measuring unit distortion and overall continuity distortion (WJ). The four features were: local phonetic context, pitch, duration and power.

To evaluate the relationship between objective cepstral and subjective perceptual distance measures, we asked subjects to score a set of utterances selected using different weights, and compared their scores with the cepstral measures. Six subjects were presented with speech synthesised by concatenation according to the different weightings.

The material consisted of seven combinations of features for selecting two sentences, each repeated three times, in randomised order (giving 45 utterances in all, including three dummy sentences to allow for acclimatisation).

Tests were administered by computer, and no records were kept of the time in each waveform where subjects detected a poorly-selected unit, only of the response counts. Subjects were asked to simply indicate perceived discontinuities ("bad segments") in each test utterance by pressing the return key, and to ignore any 'clicks' arising from simple abutting of segments. In the ideal case, where suitable segments were selected, the amount of noise was minimal because like was being joined with like. The typical length of segment was about two phones, and in the average case, a discontinuity was noticeable between one in four pairs.

In practice, perceptual scores varied considerably between respondents, with some appearing much more sensitive to abutment noise than others, but after counts were normalised per speaker, analysis of variance showed no significant effect for speaker, nor for utterance, but a clear difference for selection weighting type ($F(6,231) = 5.865$), confirming that the preferences were in agreement in spite of the differences in individual sensitivity.

The following table compares results from the perceptual test with the cepstral distances, for some range of weights. The perception measure represents the average score for each selection type, normalised by subject and target sentence. It is in standard deviation units. (PC is phonetic context)

WJ	$WU = 1.0$				Perceptual Measure	Cepstral Distance
	PC	Power	Dur	Pitch		
0.6	0.6	0.0	0.333	0.333	-0.55	0.2004
0.6	0.2	0.666	1.0	0.0	-0.49	0.2004
0.2	0.2	1.0	0.666	1.0	-0.38	0.1967
0.2	0.2	1.0	0.333	0.333	-0.07	0.1981
0.6	0.4	0.333	1.0	0.333	0.12	0.1981
0.8	0.4	0.0	0.0	0.0	0.24	0.2050
0.8	0.6	0.0	0.0	0.0	0.55	0.2056

The cepstral distance seems to give more importance to unit distortion at the expense of continuity distortion. Human perception favours more weight on WJ (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 WJ to WU. Obviously a better automatic distance measure is required which appropriately penalises burst errors. Although the numeric distances of the cepstral measure are small, the quality of the synthesis varies from very jumpy almost unrecognizable speech to undetectable unit concatenation producing natural sounding speech.

5.2 Test 2: reduced database

A further test was performed with a reduced database of Japanese. The units for the synthesiser were selected from a corpus of 503 magazine and newspaper sentence readings. In Japanese, which is not a stress-based language, there is not as great a range of prosodic variation as in English, and the damage to naturalness caused by subsequent signal processing is less, but the inclusion of prosodic information in the selection process ensured selection of more appropriate units according to an acoustic measure of spectral characteristics.

The source database consisted of 26,264 phone segments, and included 70 original labels (including segment clusters that could not be reliably separated by hand labellers). After processing, these formed 635 non-uniform units ranging in length from 1 to 7 original labels. It was pruned to a maximum depth of 35 tokens each.

Target segment labels and raw prosodic values of duration, mean pitch, and energy for each phone were extracted from a test set of 100 randomly selected sentences, and each original sentence was removed from the database before resynthesis to ensure that no original units were included. The resynthesised version was then compared with the original, using measures of cepstral similarity. Comparisons were made between the original recording of each sentence, and resynthesised versions with and without prosodic selection.

Non-weighted Euclidean measures of the cepstral distance between the original utterance and each resynthesised version were calculated on a phone-by-phone basis using 12 coefficients per 10 msec frame from LPC cepstral coding of the waveforms. Results confirmed that an improvement in spectral match was gained by inclusion of prosodic information in the selection (*seg only vs. seg+pros:* $t = 4.484, df = 6474, p < 0.001$).

Quartiles of the Euclidean cepstral distance measure

	min	25%	median	75%	max
segmental context alone:	0.0071	0.3965	0.8581	1.7219	8.8546
segmental & prosodic ctxt:	0.0073	0.3167	0.6232	1.4390	10.3748

The improved spectral match in turn confirms a strong connection between prosodic and segmental characteristics of the speech waveform, and shows that the inclusion of prosodic information in the selection of units can result in more natural sounding synthetic cpeech.

6 Discussion

We have shown that prosodic variation has more than a small effect on the spectral characteristics of speech, and that advantage can be taken of this in the selection of units for concatenative synthesis. We have also shown that a database of non-uniform units can be automatically generated from a labelled corpus and that the prosodic characteristics of contour shape and excursion can be automatically coded. Nothing above will make up for the lack of an appropriate unit in a corpus, and careful choice of this resource is essential, but a way of making better use of the supposedly redundant duplicate tokens has been suggested.

Most concatenative synthesis methods still employ a relatively small fixed number of source units, under the assumption that any modification of their inherent pitch and duration can be performed independently at a later stage through signal processing. The distortion of the synthesised speech, that is introduced as a result of changing a segment's prosodic characteristics, has until recently been masked by the generally poor, mechanical quality of the generated speech after it has passed through a coding stage. However, as synthesis quality has improved, and as the memory limitations of earlier systems are eased, it now becomes necessary to reconsider the merits of such small unit sets.

Future refinements to objective measurement procedures must include a bias to the cepstral distance measure to increase sensitivity to local concatenation points ('burst errors') and hence better approximate the human preferences. It should also be noted that such acoustic measures, because of the necessary time-alignment, are blind to inappropriate durations, and will not degrade under sub-optimal timing patterns, so for this reason too, they may not correlate well with human perception.

7 Conclusion

This paper has addressed several aspects of the concatenative synthesis process. We have shown that a large corpus of naturally-occurring speech can be used as a source of units for concatenative synthesis, and have described the tools and processes that we use for this.

The prosodic labelling is generated automatically from the speech data. The method is thus relatively language independent, and relies only on a) an adequate size corpus from which to draw the units, and b) a suitable language interface module by which to generate the transcriptions and predict the prosody for the utterance to be synthesised.

By specifying prosodic variation in terms of variance about a mean, and the slope of prosodic contours in terms of the differential of the normalised measures, we gain the advantage of speaker-independence in our synthesiser processes. The higher-level prosody prediction modules can now specify their targets in normalised terms, and whatever the database, regardless of the prediction values, the retrieved values are constrained to be within the natural range for the speaker's voice. Describing the pitch of a segment as, for example, 'moderately high and rising' ensures that the closest unit in the database will be selected, and in many cases the difference between the desired target pitch and retrieved unit's original will be small enough to be perceptually insignificant.

Our unit-generation algorithm produces a unit set that models the collocation frequencies of the input data in terms of its own labels by grouping them into equally-likely non-uniform compound units.

At the waveform level, we have questioned the validity of applying signal-processing techniques to warp the prosody of a speech segment, preferring instead to select appropriate units to minimise such post-selection distortion. We have shown simple and efficient ways to do this.

Experience with this system encourages us to believe that in the majority of cases it is better to relax our target goals in the direction of the database events rather than to impose an unnatural (and possibly distorting) pitch or duration on the waveform.

The method is currently being tested with several databases from different speakers of both English and Japanese, under different labelling conventions, and appears immune to differences in language or labelling conventions.

Acknowledgments: The authors would like to take this opportunity to thank Dr. Yasuhiro Yamazaki for supporting this research, and Drs. Yoshinori Sagisaka, Norio Higuchi, and two anonymous reviewers for comments on an earlier version of this paper.

8 References

[Bla94] A. W. Black, & P. Taylor, CHATR: a generic speech synthesis system, in Proceedings of COLING-94, Kyoto, Japan", pages 983-986, 1994,

[Bou94] O. Boeffard &, F. Violaro Improving the robustness of text-to-speech synthesizers for large prosodic variation, pp 111-114 in Proc 2nd ESCA W/S in Speech Synthesis, Mohonk, 1995.

[BU-95] M. Ostendorf, P. Price, & S. Shattuck-Hufnagel the Boston University Radio News Corpus, forthcoming

[Cam92a] W. N. Campbell. Syllable-based segmental duration. In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pages 211–224. Elsevier, 1992.

[Cam92b] W. N. Campbell. Prosodic encoding of English speech, pp 663-666 in Proc ICSLP-92, Banff, 1992.

[Cam95] W. N. Campbell & M. E. Beckman. Stress, Loudness, and Spectral Tilt, 3-4-3 in Proc Acoustical Soc. Japan, Spring meeting, 1995.

[ERL93] Entropic Research Laboratory, Inc, ESPS/waves+ , 600 Pennsylvania Avenue, Washington DC 20003.

[Gau89] J. Gauffin & J. Sundberg, Spectral correlates of glottal voice source waveform characteristics, in Journal of Speech & Hearing Research, pages 556-565, 1989.

[ISO75] International Standard ISO 532-1975(E): Acoustics - Method for calculating loudness level. 1975.

[Iwa93] N. Iwahashi, N. Kaiki, and Y. Sagisaka Speech segment selection for concatenative synthesis based on spectral distortion minimisation. Trans. IEICE vol. E76-A, 11, November 1993.

[Jon95] K. de Jong, The supraglottal articulation of prominence in English: linguistic stress as localised hyperarticulation in Journal of the Acoustical Society of America, pages 491-504, 97(1), 1995.

[Lin90] B. E. F. Lindblom Explaining phonetic variation: A sketch of the H&H theory in Speech Production and Speech Modelling edited by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pages 403-409, 1990

[Nak88] S. Nakajima & H. Hamada, Automatic generation of synthesis units based on context-oriented clustering. Proc IEEE ICASSP, pages659-662, 1988.

[Nak94] S. Nakajima, Automatic synthesis and generation of for English speech synthesis based on multi-layered context-oriented clustering. in Speech Communication, vol 14, pages 313-324, 1994.

[Rab93] L. Rabiner and B. Juang. *Fundamentals of speech recognition.* Prentice Hall, 1993.

[Sag92] 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.

[Slu93] A. M. C. Sluijter & V. J. van Heuven, Perceptual cues of linguistic stress: intensity revisited, pp 246-249 in Proc. ESCA Prosody W/S, Lund 1993.

[Slu94] A. M. C. Sluijter & V. J. van Heuven, Spectral tilt as a clue for linguistic stress, presented at 127th ASA, Cambridge, MA. 1994.

[Tak92] K. Takeda, K. Abe, and Y. Sagisaka. On the basic scheme and algorithms in non-uniform unit speech synthesis, In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pages 93-106. Elsevier, 1992.

[Tra91] H. Traunmüller, Functions and limits of the F1:F0 covariation in speech, in PERILUS XIV, Department of Phonetics, pages 125-130, Stockholm University, 1991

[Wig95] C. W. Wightman & W. N. Campbell, Improved labelling of prosodic structures, IEEE Trans. Sp. & Audio, forthcoming.