INTERSPEECH 2009 BRIGHTON

A Novel Approach to Cost Weighting in Unit Selection TTS

Jerome R. Bellegarda

Speech & Language Technologies
Apple Inc., Cupertino, California 95014, USA

jerome@apple.com

Abstract

Unit selection text-to-speech synthesis relies on multiple cost criteria, each encapsulating a different aspect of acoustic and prosodic context at any given concatenation point. For a particular set of criteria, the relative weighting of the resulting costs crucially affects final candidate ranking. Their influence is typically determined in an empirical manner (e.g., based on a limited amount of synthesized data), yielding global weights that are thus applied to all concatenations indiscriminately. This paper proposes an alternative approach, based on a data-driven framework separately optimized for each concatenation. The cost distribution in every information stream is dynamically leveraged to locally shift weight towards those characteristics that prove most discriminative at this point. An illustrative case study underscores the potential benefits of this solution.

Index Terms: concatenative speech synthesis, unit selection, candidate ranking, cost weighting.

10.21437/Interspeech.2009-252

1. Introduction

In concatenative text-to-speech (TTS) synthesis, the selection of the best unit sequence is cast as a multivariate optimization task, where the unit inventory is searched to minimize suitable cost criteria across the whole target utterance [1]. Each cost criterion encapsulates a different aspect of acoustic and prosodic context at any given concatenation point, via appropriate constraints on, e.g., inter-unit discontinuity, overall pitch contour, local duration profile, etc. [2]. Each of these constraints then leads to one of several distinct *information streams*, which collectively provide pertinent evidence to support promising candidate units. Combining that evidence then serves as the basis for final candidate ranking and selection.

Suitable information streams readily emerge from the large body of work in the literature analyzing what general features of acoustic and prosodic context most influence perception (see, e.g., [3]). To ensure that each individual cost function meaningfully scores every candidate unit relative to all others which may be potentially relevant in the given context, one can similarly rely on the many different approaches that have been proposed over the years to assess various aspects of perceptive quality (cf., e.g., [4]). When it comes to the combination of these distinct outcomes, however, the same measure of scrutiny has not been applied. Compared to the level of sophistication commonly displayed in ranking candidate units within each separate information stream, combining evidence across streams remains somewhat unprincipled.

The usual approach is to form a weighted linear combination of the various individual costs. Unfortunately, these costs are often too heterogeneous for a direct, quantitative comparison. This considerably complicates the determination of the weights, because any attempt at cross-stream normalization is necessarily burdened with technical decisions which may heavily influence any subsequent ranking and selection. Ideally, each information stream should be accounted for according to its intrinsic relevance to the optimization at hand. In practice, however, the weights are adjusted in a fairly *ad hoc* manner, in many cases by perceptually evaluating some test sentences synthesized using a limited range of weight values (see, e.g., [1]).

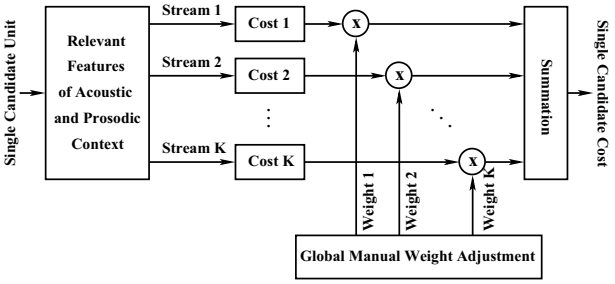

Figure 1: *Conventional Cost Weighting Framework.*

The effectiveness of this method can be debated at several levels. First, for practical reasons, the underlying synthesized material is inherently confined to a tractably small number of utterances, sometimes not even particularly representative of the eventual domain of use. Thus, it may well yield a set of weights which does not meaningfully generalize beyond the initial environment considered. In addition, this strategy makes the implicit assumption that these (globally determined) weights will apply equally well to all concatenations.

In contrast, the next section motivates a decentralized approach to multiple stream combination, under which cost weighting would be separately optimized for each concatenation considered. This promotes the novel avenue presented in Section 3, based on a data-driven framework which allows for dynamic adjustments at every concatenation point. The outcome is a scalable, fully unsupervised procedure for combining costs associated with different streams. Finally, Section 4 analyzes in detail the typical behavior of this solution via a simple but illustrative case study. These experiments underscore the potential benefits of the technique for concatenative synthesis.

2. Motivation

The conventional approach to cost weighting is depicted in Fig. 1. For each candidate unit available at a given concatenation point, multiple features are extracted to characterize relevant aspects of acoustic and prosodic context at this point. Examples of such features include, among others, the degree of discontinuity from the previous unit; the departure from ideal values for such prosodic entities as pitch, duration, and prominence; the location of the candidate unit in the recorded utterance; the spectral quality relative to the average matching unit present in the unit inventory; and so on (cf. [1] – [3]).

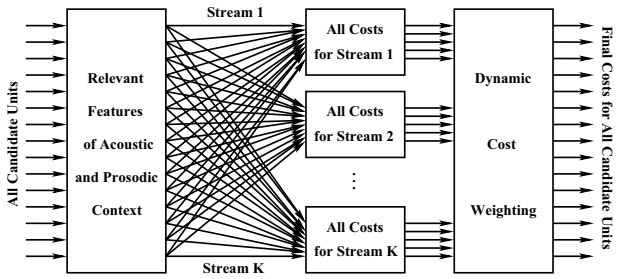

Figure 2: *Concatenation-Specific Cost Weighting Framework.*

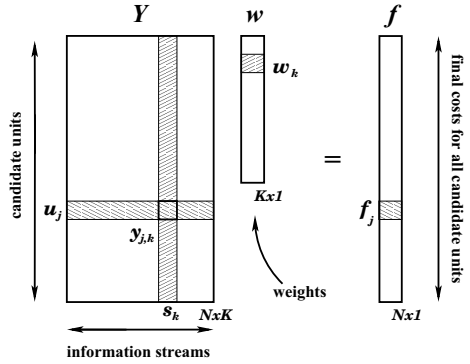

Figure 3: *Stream Weighting via Linear Combination.*

Each such feature thus gives rise to a separate information stream, within which the candidate unit can be independently assessed in terms of a specific analysis dimension. This quantitative assessment is typically rendered in the form of a non-negative penalty cost, which reflects how well the candidate scores along that particular dimension. Next, each cost is weighted using a set of global parameters, which tend to be (at least partly) adjusted on the basis of subjective human listening. Finally, the weighted costs are summed to obtain the overall cost for this candidate unit, and the unit with the lowest cost is selected as the best candidate.

This framework has a number of drawbacks, including the need for human supervision and the dependence on a necessarily small amount of synthesized material. But perhaps even more importantly, the resulting (global) weights can only be applied indiscriminately across all concatenations. Since there is no reason to believe that a single combination of costs works equally well everywhere, it would seem more desirable to reason on a per concatenation basis, as depicted in Fig. 2.

In this framework, *all* candidate units at a given concatenation point are considered simultaneously. The same features are extracted, leading to the same individual penalty costs. All costs from all candidates, however, are now gathered before any weighting is attempted. This allows, for each information stream, the cost distribution across all pertinent candidates to meaningfully inform the weighting process. A very narrow distribution, for example, suggests that the associated information stream is essentially immaterial to the ultimate ranking. In contrast, a very broad distribution is evidence of a wide range of behavior in the corresponding feature, making the stream potentially critical to the final decision. Thus the weights can be automatically adjusted based on the specifics of the available unit inventory at that point. Then all final costs are produced for all candidates at the same time, and, as before, the unit with the lowest cost is selected as the best candidate.

Several different avenues can nominally be followed to implement the framework of Fig. 2. For example, final costs could be produced via standard voting methods or other well-known learning and classification techniques. Such statistical approaches, however, require the collection of a large annotated training set, and may well involve unrealistic assumptions like stream independence. The next section proposes an alternative solution, which circumvents such difficulties by directly leveraging the candidates available in the unit inventory.

3. Proposed Solution

Consider a particular point in the synthesis process, and assume that the unit inventory contains N possible candidates at that point. Further assume that K distinct information streams are being appraised, each associated with a different aspect of perceptual quality (discontinuity, pitch, duration, etc.). Each of these streams renders an independent assessment of every candidate unit as a non-negative cost denoted by $y_{j,k}$, where $1 \leq j \leq N$ and $1 \leq k \leq K$.

The first step is to construct the $(N \times K)$ matrix Y with elements $y_{j,k}$, as illustrated in the left-hand side of Fig. 3. Each row u_j corresponds to the vector of all costs associated with an available candidate unit across all information streams, and each column s_k corresponds the vector of all costs associated with an individual information stream across all candidate units.

The overall cost for a given candidate is then computed via a weighted linear combination of all individual costs for this unit, which can be expressed as (cf. Fig. 3):

$$Y\, w \;=\; f\,, \tag{1}$$

where f is the vector of final costs f_j for all candidate units $(1 \leq j \leq N)$, and w is the (unknown) vector of desired weights w_k $(1 \leq k \leq K)$.

Since the goal is to uncover the intrinsic relevance of every information stream at the given point, the solution to (1) should presumably reflect linear combinations of the streams that correspond to directions of maximal variance in the data. This can be accomplished by finding the smallest final cost among that set of final costs f_i where individual f_i's are as uniformly large as possible, to achieve the greatest degree of discrimination between them. This is an instance of a (constrained) minimax problem, the type of which often arises in computational geometry [5]. In the present case, we can cast it as the problem of maximizing the L_2-norm of f, viz.:

$$\|f\|^2 \;=\; w^T Y^T Y w \;=\; w^T Q\, w\,, \tag{2}$$

where $Q = Y^T Y$ (with ^{T} denoting matrix transposition), subject to the (convex hull linear combination) constraints that:

$$\|w\|^2 \;=\; w^T w \;=\; 1\,, \tag{3}$$

$$w_k \geq 0\,, \qquad 1 \leq k \leq K\,, \tag{4}$$

and then selecting the component with minimal value from the resulting optimal final cost vector f^*.

Without the positivity constraint (4), the formulation (2)–(3) would amount to a standard quadratic program [5]. The requirement that the weights be all positive, however, considerably complicates the mathematical outlook. In fact, this constraint is somewhat analogous to the cardinality constraint considered in [6], which is known to make the optimization problem NP-hard and therefore intractable. We have therefore no

choice but to temporarily relax the last constraint, thereby allowing negative weighting on one or more streams.

Let us thus focus on (2)–(3) only. Clearly, the $(K \times K)$ matrix Q is real, symmetric, and positive definite, which means there exist matrices P and Λ such that:

$$Q = P \Lambda P^T , \tag{5}$$

where P is the orthornormal matrix of eigenvectors p_k (i.e., $P^T P = P P^T = I_K$, where I_K is the identity matrix of dimension K) and Λ is the diagonal matrix of eigenvalues λ_k, $1 \leq k \leq K$. In addition, the quadratic form in (2)–(3) readily leads to the associated Rayleigh-Ritz quotient $w^T Q w / w^T w$ [7]. From the Rayleigh-Ritz theorem, this quotient is known to obey the analytical bounds:

$$\lambda_{\min} \leq \frac{w^T Q\, w}{w^T w} \leq \lambda_{\max} , \tag{6}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ refer to the smallest and largest eigenvalues in Λ, respectively. Furthermore, equality at both ends of (6) is achieved with the corresponding unique eigenvector solutions $p_{\min}$ and $p_{\max}$, respectively. Hence, (2)–(3) is maximized when w is set equal to $p_{\max}$.

In the present context, this solution is not immediately admissible, since the elements of $p_{\max}$ are not, in general, non-negative. It is therefore necessary to restore the positivity constraint (4), possibly at the expense of a trade-off on the normality constraint (3). To keep things tractable, we assume that this can be done via a simple transformation of the coordinates of $p_{\max}$, such as, for example, squaring them.

Accordingly, we select the (near-)optimal weight vector w^* to be:

$$w^* = p_{\max} \cdot p_{\max} , \tag{7}$$

where the operator $\cdot$ denotes component-by-component multiplication. The ensuing final cost vector follows from (1):

$$f^* = Y\, w^* , \tag{8}$$

which in turn yields the index of the best candidate unit at the concatenation considered:

$$j^* = \arg \max_{1 \leq j \leq N} f_j^* . \tag{9}$$

Interestingly, a side benefit of this framework is that all final costs f_j^* are computed simultaneously, which makes an N-best outcome straightforward to achieve, if desired.

4. Experimental Validation

To validate the basic concept, we concentrated on a simple, but illustrative, case study, inspired by material extracted from the "Alex" male voice database deployed in MacinTalk, Apple's TTS offering on MacOS X. Qualitatively, this database is fairly similar to the Victoria corpus described in detail in [8]. In particular, recording conditions closely follow those mentioned in [8], though individual utterances generally differ.

As it turns out, one of those utterances is the sentence:

Bottom lines are much shorter. (10)

spoken in a straightforward, declarative manner. We thus opted to focus on the closely related sentence:

Bottom lines are much longer. (11)

which only differs in the last word, and is otherwise expected to have similar pitch and duration patterns as the original recorded utterance. For comparison purposes, we synthesized two different renditions of (11): one using default stream weighting (i.e., with manually adjusted global weights), and one using the dynamic weighting framework described above.

For ease of analysis, in both cases we considered only $K = 4$ information streams, namely: (i) the concatenation cost calculated between the candidate and the previous unit, (ii) the pitch cost calculated between the ideal pitch contour and that of the candidate, (iii) the duration cost calculated between the ideal duration profile and that of the candidate, and (iv) the position cost calculated between the ideal location within the utterance and that of the candidate.

Because the two sentences (10) and (11) are so close, we expected our baseline (word-based) unit selection system to assemble the first four words of (11) by pulling out the associated recorded material from (10), and only fetch the last word from elsewhere in the database. But this is not what we observed with default stream weighting. Instead, only the initial portion of (11), *bottom lines*, was picked from the original utterance (10), and the remaining material was selected from some other recordings. We conjectured that this might be a consequence of global weighting. Such would be the case, for example, if the default weights happened to be suboptimal for the last three words of the sentence.

To ascertain the matter, we next turned to dynamic weighting. For each word in the sentence, we extracted from the unit inventory all available (word) candidates, namely $N = 16$ instances for *bottom*, $N = 10$ instances for *lines*, $N = 796$ instances for *are*, $N = 92$ instances for *much*, and $N = 11$ instances for *longer*. In all five cases, we assembled the resulting $(N \times K)$ input matrix, and then computed the dynamic weights and final costs as detailed in the previous section.

This approach led to the exact same candidates being ultimately selected for the words *bottom*, *lines*, and *longer*. This time, however, different units were picked for both *are* and *much*, namely the contiguous candidates from (10) that we had originally expected to be chosen. As for the units settled on using default stream weighting, they were now relegated to ranks 15 and 17, respectively.

Closer examination of the selected candidates also revealed that, for both *are* and *much*, there were substantial differences in the overall cost profiles obtained. In the dynamic case, the contiguous candidates had a significantly lower final cost than any non-contiguous units, reflecting a much greater emphasis on the concatenation stream. On the other hand, in the default (global) case, the contiguous candidates ranked only in the middle of the top tier, suggesting that perhaps contiguity information was not given sufficient prominence.

To gain further insights into this behavior, we examined the different weight vectors (containing the $K = 4$ weights corresponding to the (i)-(iv) information streams above) generated at the various concatenation points. In the default case, of course, a single instance applied across all concatenations, given by:

$$w_0 = (0.125, 0.5, 0.25, 0.125) . \tag{12}$$

By inspection, this weight vector reflects a dominant emphasis on pitch, followed by duration, followed in equal measure by contiguity (concatenation) and position information.

In contrast, for the concatenation going into *are*, for instance, the dynamic weight vector turned out to be:

$$w_{are} = (0.98, 0, 0.02, 0) , \tag{13}$$

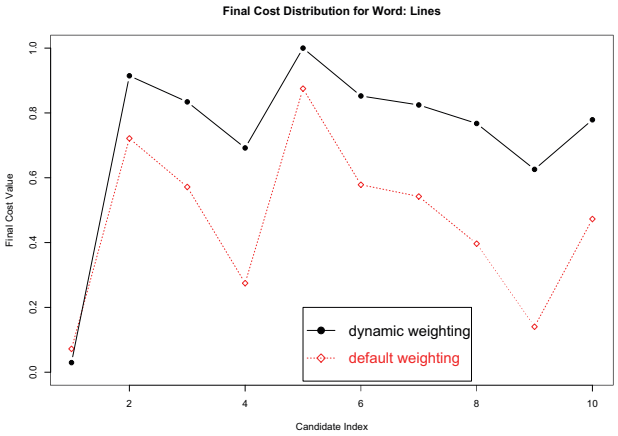

Figure 4: *Overall Cost Distributions, Concatenation → Lines.*

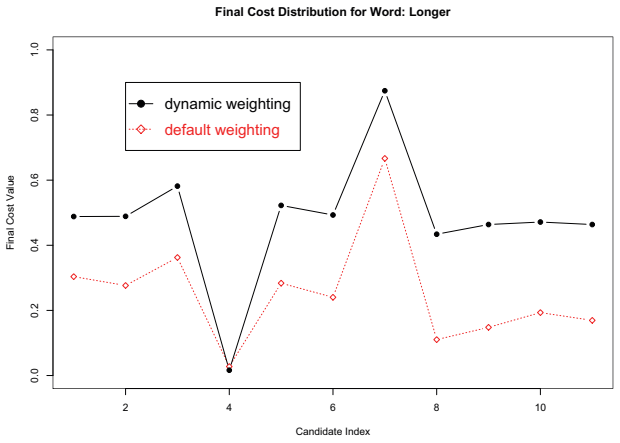

Figure 5: *Overall Cost Distributions, Concatenation → Longer.*

which overwhelmingly favors contiguity, and completely discards both pitch and position information. This seems intuitively reasonable, as for this function word co-articulation was always somewhat noticeable, while pitch contour and duration profile were both relatively similar across all available units.

Even though for three of the words the same candidates were ultimately picked, the dynamic weight vectors returned by the algorithm were markedly different as well. For the concatenation going into *lines*, for example, the default value (12) changed to:

$$w_{lines} = (0.61, 0.21, 0.18, 0)\,, \tag{14}$$

which again reflects a substantial shift toward contiguity, at the expense of the other three streams.

To illustrate the practical importance of this shift, the ensuing overall cost distribution (solid black circles) is plotted in Fig. 4 along with the baseline overall cost distribution (dashed red diamonds), across the $N = 10$ candidates available. Clearly, the new weights lead to a much better discrimination between Candidate 1 (eventually selected in both cases) and Candidate 9 (a dangerously close second in the default case, but comfortably far away in the dynamic case).

Finally, although in (13)–(14) contiguity was clearly dominant, this was not systematically the case. For the concatenation going into *longer*, for instance, the dynamic weight vector was:

$$w_{longer} = (0, 0.15, 0.15, 0.7)\,, \tag{15}$$

which this time completely discards contiguity information.[1] Instead, the most discriminative aspect now becomes the position within the utterance. This again makes a great deal of intuitive sense, given the decisive impact of pre-pausal lengthening on the last word of the sentence.

Fig. 5 compares the ensuing overall cost distributions across the $N = 11$ candidates available, in a manner analogous to Fig. 4. With the new weights, there is a much better discrimination between Candidate 4 (eventually selected in both cases) and Candidate 8 (as above, dangerously close in the default case, but a considerably more distant second in the dynamic case).

[1] Since no contiguous unit was available at this point, this suggests that all available candidates happened to have comparable concatenation costs, which was indeed verified empirically.

5. Conclusion

We have proposed a novel strategy for combining the various costs encapsulating different aspects of acoustic and prosodic context in unit selection TTS. This approach leverages the cost distribution in each information stream in order to dynamically determine, on a per concatenation basis, the relative importance of every stream to overall unit ranking. Unlike conventional methods, this framework thus results in a cost weighting solution which is separately optimized for each concatenation considered. This in turn allows candidate units to be ranked according to those characteristics that are intrinsically most discriminative at each point.

Ensuing benefits have been objectively characterized in terms of the ability of the TTS system to select more contiguous units for a given pitch and duration profile, and more generally to emphasize one information stream over another in an intuitively reasonable manner. We are currently in the process of conducting subjective evaluations to confirm the perceptual advantages of this solution.

6. References

[1] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using Large Speech Database," in *Proc. ICASSP*, Atlanta, GA, pp. 373–376, 1996.

[2] W.N. Campbell and A. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis," in *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Hirschberg, and J. Olive, Eds., New York, NY: Springer–Verlag, pp. 279–292, 1997.

[3] T. Dutoit, *An Introduction to Text–to–Speech Synthesis*, Norwell, MA: Kluwer, 1997.

[4] E. Klabbers and R. Veldhuis, "Reducing Audible Spectral Discontinuities," *IEEE Trans. SAP*, Vol. 9, No. 1, pp. 39–51, January 2001.

[5] J. Nocedal and S.J. Wright, *Numerical Optimization*, 2nd Ed., New York: Springer–Verlag, 2006.

[6] B. Moghaddam, Y. Weiss, and S. Avidan, "Spectral Bounds for Sparse PCA: Exact and Greedy Algorithms," *Adv. Neural Inform. Process. Systems (NIPS 18)*, Cambridge: MIT Press, 2006.

[7] A.W. Leissa, "The historical bases of the Rayleigh and Ritz methods," *J. Sound and Vibration*, Vol. 287, pp. 961-978, 2005.

[8] J.R. Bellegarda, K.E.A. Silverman, K.A. Lenzo, and V. Anderson, "Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation," *IEEE Trans. SAP*, Vol. SAP–9, No. 1, pp. 52–66, January 2001.