

On the Design of Cost Functions for Unit-Selection Speech Synthesis

Francisco Campillo Díaz and Eduardo R. Banga

Signal Theory Group (GTS)
Dpto. Teoría de la Señal y Comunicaciones
Universidad de Vigo (Spain)
{campillo, erbanga}@gts.tsc.uvigo.es

Abstract

The quality of the synthetic speech provided by concatenative speech systems depends heavily on the capability of accurately modeling the different characteristics of speech segments. Moreover, the relative significance or weighting of each feature in the unit selection process is a key point in the relationship between synthetic speech and human perception. In this paper we propose a new method for optimizing these weights, making a separate training according to the nature of the different parts of the cost function, i.e., the features referred to the phonetic context of the units and the features related to their prosodic characteristics. This work is mainly focused on the target cost function.

1. Introduction

Unit selection speech systems have proved to be a good choice for the generation of quite natural utterances. The existence of different instances of the same unit, demiphones in our case, provides the possibility of selecting the closest realization to the desired target, taking into account a set of phonetic and prosodic constraints. This fact, however, introduces the problem of finding a suitable group of features that properly characterizes a speech segment. Moreover, the relative importance of such features is a key point in the selection process, and it should be directly related to human perception. This way, the weight of the fundamental frequency or the spectral continuity should reflect their subjective impact on the final quality of the synthetic speech. In [1] and [5], two automatic schemes for optimizing these weights are described. The first scheme considers an objective measure (the mel-cepstral distance between synthesized and natural utterances) and tries to minimize the cost function by an exhaustive search in the space of weights. The second method tries to predict the output of the objective cost function by a linear weighting of the target sub-costs. In [3], a different approach is described that tries to train the unit searching algorithm with the results of perceptual tests in an attempt to make the algorithm behave as human listeners do. Notwithstanding, these methods are not utterly complete, as they do not take into account the contribution of all the features in the process of optimization, and some of them must be adjusted by hand-tuning.

In this article we propose a novel method for training the weights of the target cost function, taking into account all the contributions according to its nature and the way they seem to affect human perception. All the quality tests in this article are based on informal listenings by people accustomed to working with text-to-speech systems and, although more rigorous tests

are necessary, we think they provide a good indicator of the improvements.

The outline is as follows: in section 2 we will describe the cost functions, mentioning the different features usually considered. In section 3 we present our weight optimizing procedure compared to other approaches. Finally, in section 4 we present some conclusions and give some guidelines for further research.

2. The cost functions

Traditionally, two cost functions are employed for finding the best sequence of units for synthesizing a desired utterance: the target and the concatenation cost functions. The former evaluates the similarity between the target and the candidate units, whereas the latter measures the cost of concatenating two candidate units. In this work we will focus mainly on the target function.

The target cost function tries to evaluate the similarity between the desired objective feature vector and the corresponding feature vector of each of the candidates. In general, apart from other factors such as the accent or the type of proposition, these features can be divided into two categories: the phonetic context and the prosodic attributes (f_0 , power and duration).

Regarding the phonetic context, a fixed-sized phoneme window (typically with a length of 3 or 5 phonemes) centered at the current unit is usually considered. In [1], the preceding and following phonemes are characterized by a vector of distinctive features (point of articulation, vowel height, consonant type...) and a comparison function is proposed that is based on equally penalizing the differences in the components of the target and candidate vectors, with a partial sub-cost of one if the components differ, and zero if they are the same. In [3], if the surrounding phonemes of the target belong to the same class that the phonemes around the candidate, the sub-cost value is set to zero and one otherwise. In previous works [6] [7], we followed a similar strategy, with a window of five phonemes centered at the current unit, and adding small sub-costs for each of the phonemes that did not exactly match. This approximation was too rude, as it penalized in the same manner very different phonetic contexts.

In order to solve this problem, in this contribution we propose the use of the centroids of the average mel-cepstral vectors of the occurrences of each demiphone. Therefore, the phonetic cost turns into the Euclidean distance between the

centroids of the surrounding demiphones of the target and the candidate units. This way, we avoid the use of a discrete distance function, whose numerical values are not directly related to the different features of the compared demiphones. Moreover, the distance between the centroid of the target demiphone and the mean cepstral vector of the candidate unit provides a good method to detect and penalize speech units that were probably incorrectly labeled or even poor realizations of sounds. Informal tests have shown a clear improvement of the quality of the synthetic speech.

With reference to the prosodic sub-costs, similar features are employed in most of the systems. For instance, in [3] the mean fundamental frequency, the slope in the intonation contour and the duration are considered. In [2], duration, log power and mean F_0 are mentioned, among other factors. In our case, instead of using the mean frequency and the f_0 slope, we consider the pitch values at the beginning and at the end of the speech unit, in order to favor the selection of a unit with the desired pitch at the points of concatenation.

Finally, we also consider some other markers, such as the binaries begin/end of word, referred to the position of the demiphone in the word, or the type of proposition.

With respect to the concatenation cost, continuity of frequency, power and spectral envelope are almost always considered, with changes in the way they are represented. In [2] the spectral envelope is parameterized by 16 mel-cepstrum coefficients and for each pair of units a search for the best concatenation point is performed. On the other hand, in [3] the normalized Euclidean distance of the first three formants is employed. In our system, we compute the Euclidean distance between the mel-cepstral vectors of the two units at the concatenation point, as it seem to have shown a good correlation with perceptual measures [4].

3. Optimizing the weights

As we have already mentioned, most approaches try to estimate an objective quality function by considering the different contributions (subcosts) to the cost functions. In [3], this quality function is substituted by perceptual scores, but it is only applied to isolated single-syllable words, and its extension to whole sentences is not trivial, as the influence of features such as f_0 or duration is stronger in the latter case. Thus, we decided to pay more attention to the two schemes described in [1].

The first method, called weight space search, looks for the set of weights that produces better results, according to the objective cost function. The problem of this approach is that the computational requirements grow exponentially with the number of weights, what leads to consider a gross partition of the possible values of each weight. There have been interesting works, such as [5], where they employ strategies like, for example, precalculating the unweighted sub-costs for each candidate, or running the unit selection process for all weights combinations before the synthesis step. The method offers a more efficient calculation, but the computational load is still a problem.

The other method, regression training, tries to approximate the values of the objective cost function by the

sub-costs of the cost functions. It has the interesting advantage of permitting a separate training for the weights of the target and concatenation cost functions, and allows considering different sets of weights for each type of phoneme.

In the next sections we will talk about the quality of the considered objective cost function, and the principal contribution of this work: a new scheme for training the weights of the target cost function.

3.1. The objective cost function

The quality of the synthetic speech obtained by using the weights computed with the previous two methods is highly dependent on the relationship between the objective cost function and human perception. In [1], the cepstral distance between original and synthetic sentences is used. Based on this work, and with the goal of analyzing the relationship of the cepstral distance with the quality of the synthetic speech, we have developed a study of the spectral distance of the units of the same type in our speech corpus, computing the Euclidean cepstral distance between all the realizations of the different demiphones. The results were organized in several groups, some of them referred to the type of phoneme (according to the classification: vowels, silence, fricatives, laterals, nasals, vibrants, voiced plosives and unvoiced plosives), and some others referred to the demiphone itself.

The considered Galician corpus consists of two parts: a set of 800 phrases designed manually to be a rich prosodic sample of the Galician language, and 500 more which were collected automatically to take into account some additional complex prosodic structures and balanced the frequency of occurrence of the demiphones in Galician. In summary, our speech corpus consist of about one and a half hours of recorded speech, and contains 120.000 realizations of demiphones, which were clustered into 970 groups according to the allophone identity with the left or right context, the type of intonational group (declarative, interrogative, exclamation,...), the position of the accent group within the phonic group (initial, intermediate and final) and the stress. The number of occurrences of each unit varies from 600-1000 for the most common demiphones in Galician, to 10-50 for the least frequent units.

First, we carried out informal listenings with the cost functions adjusted by hand-tuning. The resulting speech quality was quite good with smooth transitions between units. Soon afterward, we resynthesized the corpus sentences with the restriction that a unit could not be selected for generating the speech segment from where it was obtained. Then, we computed the cepstral distance from each selected unit to the original segment and compared its value with the precomputed distances between every pair of realizations of that demiphone.

Figure 1 shows a histogram that reflects the percentage of demiphones susceptible of being chosen with a closer distance to the original than the actually selected. The mean value is 37 %. As it was expected, most of the units are very close to the original realization. Some large distances are a consequence of the comparison method, as some demiphones

only have very few occurrences in the corpus. For example, the lateral demiphone /l+l/ (/l/ with /l/ as right context) has only four instances, so if there were two units closer than the selected demiphone, the percentage would be of 66%. These are quite rare cases (almost irrelevant) and, nevertheless, there is a non-negligible amount of units near the 100% (i.e. the worst case according only to the cepstral distance). This is an expected result since we know that there are some speech features that the Euclidean cepstral distance does not take into account (f_0 , for example). In [4], it is shown that the combination of different cepstral distances leads to a higher correlation with human perception, but it is only used for detecting spectral discontinuities, although it could be considered as a future line for research. In [8] an acoustic vector which includes mel-cepstrum coefficients, F_0 , power and delta-cepstrum is defined, but only for purposes of clustering, and not as a quality measure.

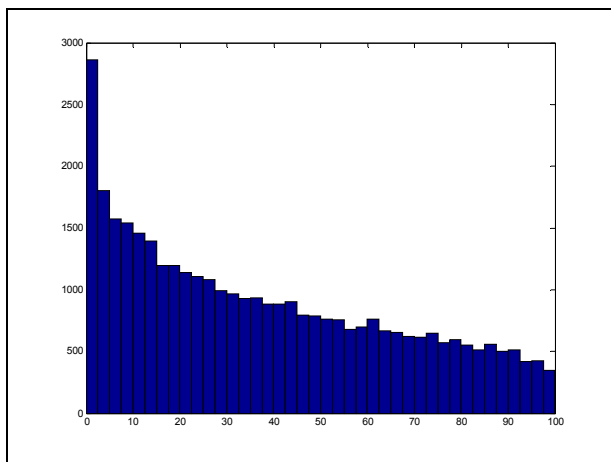


Figure 1 Percentage of units closer to original

3.2. Training the weights

With the previous limitation of the cepstral distance in mind, we tried a separated training of two sets of weights of the target cost function. The first set refers to the phonetic features, while the second set refers to the prosodic attributes.

For training the first set, we employed a linear regression method similar to the one described in [1], using the Euclidean distance of the mel-cepstrum parameters as the objective cost function. A phonetic vector with the features vowel/consonant, voiced/unvoiced and type of phoneme was assigned to each one of the four phonemes surrounding the current demiphone. The comparison function added a sub-cost of one for every distinctive component in the feature vector, and zero otherwise. The linear model obtained did not provide acceptable results, as the input information was not too related to the cepstral distance. In fact, the coefficient of explained variance was not much better than the value obtained when comparing not the phonetic vector, but the phonemes. This was an expected result, as the comparison of the different phonetic features of two phonemes can only give a qualitative idea of the differences between them, but not how close they are.

Figure 2 shows a histogram of the cepstral distance between every pair of vowels in the speech corpus, where the zero values correspond to the self-comparisons of the speech units, while the largest values are mainly due to poor realizations. Similar histograms were obtained for the other phonetic classes. We have also compared the histograms of the distance between all the occurrences of some demiphones, and the results were alike.

As it was presumed, the cepstral distances between demiphones of different type were larger. This fact led us to substitute the sub-cost associated with the phonetic context for the summation of the Euclidean distances between the cepstral centroids of the surrounding demiphones. The linear model improved substantially with the new cost functions, and informal listenings showed a better performance. The coefficient of explained variance was about 30%, a good result if we consider that we are trying to model the distance between two demiphones as a function of the distances between the centroids of the surrounding demiphones. In addition, as a step by step linear regression is employed, an idea of the significance is obtained, that reveals the adequacy of the sub-costs related to cepstral distances, in contrast with the former values given by the qualitative sub-costs.

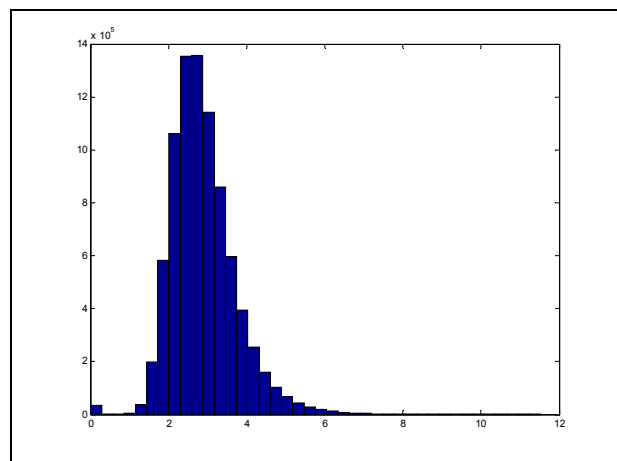


Figure 2 Vowel Histogram

Regarding the weights of the prosodic cost, i.e., frequency and duration (our system does not currently consider energy as a parameter of the target cost function) are treated in a very different way. These sub-costs should be dependent on the amount of distortion introduced in the process of prosodic modification. For example, our system does not modify those units whose prosodic features are close enough to the target requirements in order to reduce the inherent distortion of this kind of process. Therefore, in such situations the prosodic sub-cost should be zero

Thus, we consider two thresholds for each prosodic sub-cost, \mathbf{Th}_{\min} and \mathbf{Th}_{\max} . If the difference between the target and the candidate feature (f_0 or duration) is below \mathbf{Th}_{\min} , that feature is not modified and the corresponding sub-cost is set to zero. On the contrary, if the difference is greater than \mathbf{Th}_{\max} , the sub-cost will take its maximum value, in order to try to discard that candidate and select some other unit available. Between these two thresholds, the sub-cost function

is considered linear as illustrated in Figure 3. Currently, we are using 5 and 20 Hertz as the thresholds for fundamental frequency, and 20 and 40 milliseconds for duration

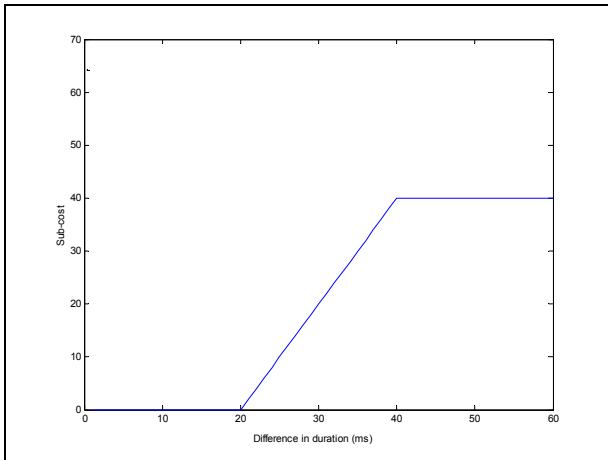


Figure 3 Example of duration sub-cost

In this manner, the target cost function is equal to the sum of the phonetic cost and the prosodic cost. Both contributions must be balanced with the α coefficient shown in expression 1. This single weight can be adjusted by a subjective test without a great effort, although further research in this direction is under development.

$$C_{target} = \alpha * C_{phon} + (1 - \alpha) * C_{pros} \quad (1)$$

With regards to the concatenation cost function, its weights are not optimized yet in the present version of our synthesizer. However, they were adjusted by hand-tuning, and we think the performance is quite reasonable.

4. Conclusions

The optimization of the weights of the cost functions is a fundamental point in the quality of the synthetic voice of a unit-selection speech synthesis. The weights reflect the relative importance of each feature, and serve as a mapping between the differences in each feature and the differences in human perception.

Many systems use the cepstral distance as a measure of the quality of the synthetic voice. However, some features of the speech segments are not taken into account by this type of function (f0, for example). To solve this problem, we consider the different nature of the features of the target cost function. As most of the systems do not modify the spectral envelope and, nevertheless, perform prosodic modification, we decided to use cepstral distance to train the weights related to the phonetic context. However, the weights related to frequency and duration are adjusted to favor none or slight manipulations in order to reduce additional distortions. Finally, we propose to balance the phonetic and prosodic costs by means of a single weight.

In the next future, we will also focus our work in the optimization of the concatenation cost function. Finally, as

our text-to-speech system also employs unit selection for intonation generation [7], we will try the optimization of the related cost functions as well.

In <http://www.gts.tsc.uvigo.es/cotovia> there is an online demo of our Galician text-to-speech system (and a preliminary Spanish version).

5. Acknowledgements

This work has been partially supported by the “Ministerio de Ciencia y Tecnología”, FEDER funds and the “Xunta de Galicia” under the projects TIC2002-02208, PGIDT01PXI32205PN and PGIDT02PXI32201PR.

6. References

- [1] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using large speech database”, ICASSP96, V1, pp 373-376.
- [2] A. Black and N. Campbell, “Optimizing selection of units from speech databases for concatenative synthesis”, Eurospeech95, pp 581-584.
- [3] M. Lee, D. P. Lopresty and J.P Olive, “A tex-to-speech platform for variable length optimal unit searching using perceptual cost functions”, Proceedings of the Fourth ISCA workshop on speech synthesis, August-September 2001, Perthshire, Scotland, pp. 75-80
- [4] J. Vepa, S. King and P. Taylor, “New objective distances measures for spectral discontinuities in concatenative speech synthesis”, Proceedings of the IEEE workshop on speech synthesis, 2002, Santa Monica, California, USA.
- [5] Y. Meron, K. Hirose, “Efficient weight training for selection based synthesis”, Eurospeech99, V5, pp. 2319-2222.
- [6] E. R. Banga, F. Campillo, E. Fernández and F. Méndez, “Sistema de conversión texto-voz en lengua gallega basado en la selección combinada de unidades acústicas y prosódicas”, Procesamiento del Lenguaje Natural, Revista nº 29, pp. 153-158.
- [7] F. Campillo and E. R. Banga, “Combined prosody and candidate unit selections for corpus-based text-to-speech systems”, Proceedings of ICSLP02, pp.141-144.
- [8] A. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis”, Proceedings of Eurospeech97, V. 2, pp. 601-604.