Linguistic features weighting for a Text-To-Speech system without prosody model

Vincent Colotte[1], *Richard Beaufort*[2]

[1]LORIA - Speech team, University Henri Poincare, Nancy, France
[2]Speech department, TTS group, Multitel ASBL, Mons, Belgium
colotte@loria.fr, beaufort@multitel.be

Abstract

This paper presents a Non-Uniform Units selection-based Text-To-Speech synthesizer. Nowadays, systems use prosodic models that do not allow the prosody to vary as far as we should hope, involving a listening comfort degradation. Our system has the advantage to avoid the using of prosodic model. Speech units selection builds its features set exclusively from the linguistic information generated by the natural language analysis. We also present an original method to automatically weight these features. Therefore, selected units are not restricted by a predetermined prosody. With only using linguistic features, we obtain a various prosody and the units concatenation is performed without resort to heavy signal processing.

1. Introduction

Nowadays, Text-to-Speech synthesis systems are based on a sequential and modular architecture, often divided in three main step: natural language processing (NLP), units selection and digital signal processing.

At the units selection level, this resulted in a search for best fitting units in terms of language analysis and concatenation in order to avoid signal distortion. As a consequence, units selection relies on two cost parameters: a target cost giving the distance between a target unit and units coming from the corpus, and a concatenation cost estimating the acoustic distance between units to be concatenated.

All of the units are represented by a limited number of relevant features. ”Relevant” means that they provide a good representation of acoustic variations within a unit. Every system uses linguistic, acoustic and symbolic features in variable proportions. Linguistic features are directly found out by analyzing the text. Acoustic and symbolic features are predicted by prosodic models. Among the acoustic features the fundamental frequency and the duration are the most often used, while the tone is the most recurring symbolic feature.

One of the crucial point consists in the assessment, in terms of importance, of the different features being part of one or another cost. Indeed most features may not be considered as equally important; some of them are more influential on the quality of results. Consequently, some researches have been achieved to find out what could be the ideal weighting for the selection process. Among the systems which were worked out, none of them suggests an automatic weighting for all features. Besides, weighting is still implemented manually at one or another step.

The first suggested weighting, carried out by the CHATR system [1, 2, 3], implies to form a network between all sounds of the corpus. Then, a learning phase is aimed at improving the acoustic similarity between a reference sentence and the signal given by the system, by tuning the features weighting (either by successive iterations or by linear regression). There are two inherent drawbacks in this method: on the one hand, the computation load, still consuming resources (even if done *off-line*), and on the other hand, the restricted amount of features the process can weight. Most of the time, part of the weighting must be done manually. In order to reduce the computation load, [4, 5] carry out a *clustering* of sounds, prune the clusters and directly use these clusters during the selection step.

Another weighting method relies on a corpus representation achieved by a phonetic and phonologic tree [6, 7]. During the selection, [6] look for candidate units with the same context as the target unit. However, the features they used are not automatically weighted. [7] have a more ”top-down” approach which tries to find the longest possible unit among all corpus sentences. The search starts at the sentence level, looking for the sentence itself. If it can not be found, the system goes down (groups of words, words, syllables, . . .) still a unit, the longest possible one, is found in the corpus. Unfortunately, this method could suggest a word by word synthesis, which involves prosodic cuts that can disturb the listener.

Because acoustic and symbolic features are used during selection, the building step of a prosodic model seems to be necessary. Moreover, the advantage of current prosodic models, whatever they are rule- or corpus-based, relies in the correctness of their suggested standard prosody.

However, these models present few prosodic variations: same prosodic patterns are repeated sentences after sentences, which leads to a decrease of the listener’s satisfaction while listening synthesized speech. Moreover, the prosodic model is language dependent.

Among the few works that attempt to free themselves from the prosodic model, [8] uses only linguistic features for units selection: the name of the phoneme, its position into the word and its position in the syllable. Unfortunately, units selected by means of these few criteria show acoustic discontinuities, and require some signal processing. As a result, generated speech shows a less natural character.

The goal of this paper is to propose a Non-Uniform Units-based speech synthesis system, on the one hand, freed from any prosodic model and any heavy signal processing, and on the other hand, associated with a automatic method to weight the linguistic features used for the units selection. At first, we give the linguistic features extracted and used in the system. In the second and the third section, the method of linguistic features weighting and the selection step are explained. And before the conclusion, we present an evaluation of our system.

2. The LiONS system

Our system (LInguistically-Oriented Non-uniform units Selector) follows the main steps of a NUU selection system. The analysis of the text and the selection of units in a corpus. Unlike other TTS systems, our system does not use a prosodic model. Therefore, it involves two significant changes in the structure. The first one is that there is no prediction of the duration or the F0 value of the unit which should be selected. We need to exploit the only information of the NLP module: linguistic features without symbolic features as tones. And the second one is that, without features as tones which summarize prosodic behavior, we need to use a great number of features. So we have developed a automatic method to weight each linguistic feature. These weights will be used during the units selection to compute the target cost.

To obviously keep homogeneity of the system, the same linguistic analysis engine is used for the linguistic features extraction during the training and the run-time phase. This NLP comes from the NLP module of *eLite* (*"Enhanced LInguistically-based TExt-to-speech synthesizer"*), speech synthesis system developed at Multitel ASBL.

For the training and the selection, we used linguistic features as phonemic context (around the target), syllabification, kind of the syllable (CV, CVC, V, . . .), phoneme position in the syllable, number of syllables in the word, syllable position in the word, word position and number (in the sentence), etc. Of course, selection based only on linguistic information needs extra features:

- *Primary* and *secondary stress*, which are strictly linguistic information that may be extracted from phonetization lexicons,
- *Rhythm groups*, which include several types and allow to determine implicitly the position where group emphases may appear. The groups boundaries always correspond to syllabic boundaries, which come from the syllabification results. It also is linguistic information.

3. Weighting of linguistic features.

During the training step (off-line), as other systems, we need to rank features and their weighting. Unlike, some other systems, we does not built a CART to prepare a pre-selection for the units selection. We just compute the weight of each feature (that reflects its relevance) and no clustering will be kept for the pre-selection. The weights will directly be used in a traditional target cost formula in the selection step.

Because of articulatory differences, each phoneme behaves differently than the others in a same elocution context. That is the reason why a single weighting of linguistic features for all phonemes together would not be relevant; it would be better to weight these features independently for each phoneme.

For a particular phoneme, we take all its acoustic representations in the spoken corpus. By using the K-Means algorithm, these representations are split into different clusters. A cluster gathers similar acoustic representations together. In this case, the Kullback-Leibler distance [9] is used as similarity index. The initialization is set at several clusters whose acoustic representations are distributed according to their duration. There are seven duration clusters built according to the mean and the standard deviation. For each of them, the optimal number of clusters is automatically computed by maximizing the variances ratio.

Once these clusters are defined, the linguistic features weighting may start. It aims at determining how much each feature can distinguish several clusters. Each cluster can be considered as a class. Therefore, the most appropriate method to solve this problem is a *decision tree*. We can note that we will use decision tree technique not to build clusters but from the clusters previously computed.

Decision tree building relies on the concept of entropy. Entropy measures system disorder: the more disorder, the less information. Consequently, entropy computation for a list of features allows classifying them according to their intrinsic information. The more a feature has a low entropy, the more it is informative and relevant.

In our case, the entropy of feature is computed as a Gain Ratio, i.e. the ratio of Information Gain to the Split Information. For each feature, we compute its Gain Ratio according to the clusters built by K-Means. It will allow determining the features ranking between all levels of the decision tree (IGTree). From these values, we obtain for each feature the weights which will used inside the target cost computation (see section 4.1). The weights are obtained by a logarithmic scaling of the Gain Ration :

$$W_k^l = 2 \times log(10 \times GR(k)) \qquad (1)$$

where W_k^l is the weight for the feature k of the phoneme l, $GR(k)$ is the Gain Ration computed for the feature k.

However, we are not interested in the tree. Thus, the tree is not actually built. Moreover, clusters are discarded too. We just kept the value of weights which are computed from the entropy, to give the rank and the relevance of the features. This is a great difference with other systems: we do not perform preselection of candidates by using neither the clusters nor the tree. We just use features weights to compute the target cost of each candidate (see Section 4.1).

Corpus. The training of weights on the corpus which will be used for the unit selection step have the advantage to keep the consistency with the speech features of the speaker. It's not the case with usual prosodic models which are learned with several speakers or from rules to obtain a standard prosody.

Sentences of the written corpus are analyzed by eLite and labeled as seen in the previous section: number of words, syllabification, phonetization, and articulatory features of phonemic contexts for each phoneme. . .

Sentences of the speech corpus are segmented into phonemes and diphones. For each of them, acoustic features are computed: fundamental frequency, Linear Predictive Coding (LPC) coefficients and intensity. These features only take part in the concatenation cost computation (and in the previous acoustic clustering).

The labeling and our method of weighting is fully automatic.

4. Speech units selection

The speech unit selection is preceded by the linguistic analysis of the text. Like the training and the corpus annotation, we used the same NLP engine (*eLite*). For a sentence, the linguistic analysis generates the corresponding phonemes and the list of linguistic features associated to each of them. We define as a *target* every pair {*phoneme, features*}.

We underline that the analysis does not generate any symbolic feature nor predict any acoustic feature (duration or F0 value) because there is no prosodic model. In other words, we have not used higher level prosodic feature as tones. We just used the linguistic features as listed in section 2. These features cover local and global context (position in the sentence. . .) to hope to cover all the necessary information to produce a correct prosody by the only selection.

4.1. Selection step

Selection occurs in three steps: (1) pre-selection of candidates which are phonemic units, and target cost computation for each candidate, (2) transformation into a diphonic representation, and (3) selection of diphonic units minimizing the double cost $\{$*target, concatenation*$\}$.

Pre-selection. For a given target, all the candidates must at least have the same phonemic label, *i.e.* the name of the phoneme. It's the only pre-selection that we made. The target cost computation of each candidate unit is carried out at this stage. In this computation, features are weighted using the weights determined during the training. The target cost of a candidate j for a target i corresponds to a weighted summation of the "difference" between the features of the candidate and the features of the target:

$$TC(j,i) = \sum_{k=1}^{N} W_k^l \times (1 - \delta(C_k^j, T_k^i)) \qquad (2)$$

where: $TC(j,i)$ is the target cost of candidate j for target i; k varies from 1 to N, the number of features; C_k^j is the value of the feature k for candidate j; T_k^i is the value of the feature k for target i; $\delta(.,.)$ returns 1 if values are the same, 0 otherwise. W_k^l is the weight computed during the training step for the feature k for phoneme l where l is the phonemic label of the target i (see Section 3).

Diphonic representation. At this step, diphones to be selected are only those that can be formed from adjacent phonemic candidates in the corpus. However, if a target diphone does not have a candidate, we create candidates containing the target phonemes partly left or partly right-hand side, according to the diphone which one needs. The target cost of each diphonic candidate is the sum of the target costs of the two phonemic candidates which constitute it.

Unit selection. The selection is operated in a traditional way, by resolution of the lattice of possibilities using the Viterbi algorithm. The result of the selection is the path, in the lattice of diphones, which minimizes the double cost $\{$*target, concatenation*$\}$. The target cost was computed at the previous step. The concatenation cost is solved when running through the lattice of possibilities. The concatenation cost has been defined as *the acoustic distance between the units to be concatenated.* To calculate this distance, the system needs acoustic features, taken at the boundaries of the units to be concatenated: fundamental frequency, spectrum, energy and duration. The distance, and thus the cost, is obtained by adding up:

- the difference between the fundamental frequency values,
- the spectral distance (*Kullback-Leibler* type),
- the difference of energy values,
- the difference of duration of the phonemes which are used as concatenation point. For instance, if the system has to concatenate target diphones /pa/ and /aR/, it will try to favor the couple of candidate diphones whose the half /a/ (at left and at right) come from an /a/ with more or less the same duration (parent and garden, if we suppose that the /a/ have a similar duration for these words).

Of course, the sum is weighted, but the weighting, unlike the target cost, is not learned automatically during the training time: it is manually given, and favors mainly the spectral distance and the difference in fundamental frequency.

It still should be noted that the double cost $\{$*target, concatenation*$\}$ itself is weighted, so that the target cost and the concatenation cost do not have the same weight in the choice of the best candidates. Currently, this weighting is still partially manual: thus the system still relies on two dials whose adjustment is related to a corpus of some sentences to assess the quality of the system by a listener.

4.2. Speech units concatenation.

Except the concatenation itself, no signal processing is necessary. The selected diphones sequence is concatenated acoustically, using a traditional technique (*OverLap and Add* type): pitch values are used to improve the joint of diphones.

5. Evaluation

The corpus we used can be considered poorly adapted: 75 minutes of speech while a NUU corpus should contain 3 hours at least. The corpus gathers 800 sentences extracted from French broadcast news, but the female speaker has light Swiss intonation, and realizes strong prosodic variations.

However, we made an evaluation of the system. 50 subjects listened to 25 French sentences. Among them, 20 sentences were synthesized by LiONS, and 5 directly came from the corpus used for the selection. Aims of the evaluation were:
− Evaluating the *intelligibility*, the *naturalness of the prosody*, the *quality of the concatenation* and the *listening comfort*. Each criterion had a range of values from 1 to 5 (5 = the best score).
− Evaluating the distance between *synthetic* and *original voices*.

Explanation given to the subjects and sentences of the evaluation can be found on the web page http://www.multitel.be/TTS/LiONS/evaluation.html.

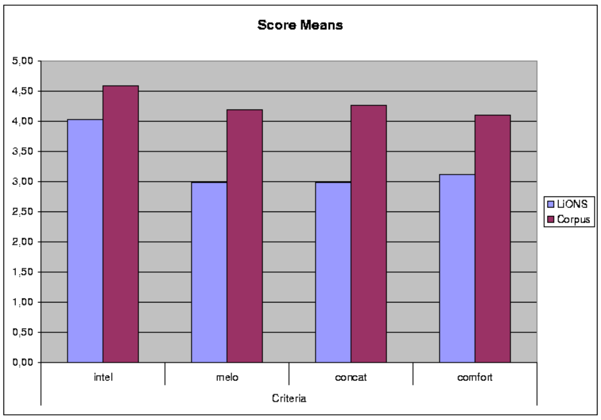

Figure 1: Score means of sentences generated by LiONS and extracted from the Corpus. Evaluating the *intelligibility* ("intel"), the *naturalness of the prosody* ("melo"), the *quality of the concatenation* ("concat") and the *listening comfort* ("comfort").

Results of the evaluation are shown in figures 1, 2 and 3. "Corpus" label is the results obtained for the 5 sentences extracted from the corpus and "LiONS" label is for the 20 sentences synthesized by our system. The general evaluation can be considered as positive. Concatenation, melody and listening comfort are felt as normal, while the intelligibility of the speech is very highlighted. The results for corpus sentences are interesting: they show the quality lack (or the subjective quality lack) of the original female voice. These sentences are less appreciated than we could expect from a human voice. We may hope better results as soon as we will have a better corpus with respect to the speaker voice quality. And therefore, we will be able to make a comparative test with other systems with prosodic model.

We give at the figure 4 the results of the selection for one of the twenty generated sentences (4th/20).

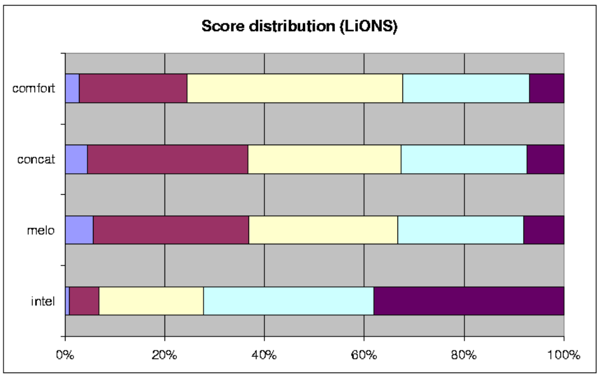

Figure 2: Distribution of scores for sentences generated with LiONS (in order 1-, 2-, 3-, 4- and 5-score).

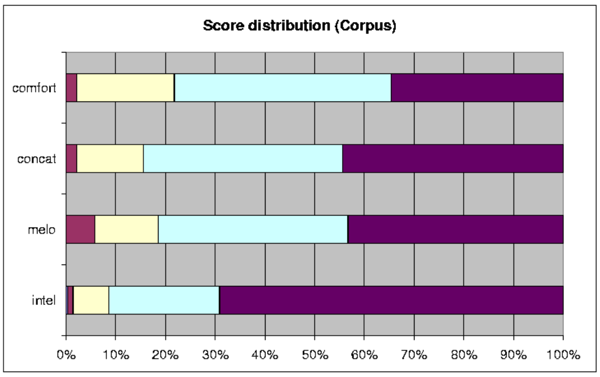

Figure 3: Distribution of scores for sentences extracted from the selection corpus (in order (1-), 2-, 3-, 4- and 5-score).

6. Conclusion

Our synthesis system is freed from any prosodic model, whatever it is acoustic or symbolic, so as to allow more variations in the prosody of the successively generated sentences. At the same time, there was the constraint to keep the advantage of traditional systems, namely their weak resort to the signal processing at the units boundaries.

To achieve this, features of the selection are chosen exclusively among linguistic information coming from the language analysis. Beside traditional linguistic criteria, such as the phonemes of context, new information has been added, like the rhythm groups, able to describe the potential prosodic behavior of units. This contribution will take more scope with the development of tools to automatically build corpus. At this moment, our database is composed of 1h15 of speech with about 800 sentences extracted from French broadcast news.

During the selection step, a target cost and a concatenation cost are computed. Features of the target cost have been automatically weighted during the training step: for each phoneme, each feature is weighted by computing its Gain Ratio from clusters of similar acoustic realizations.

Selected linguistic features and their weighting have proven their effectiveness: selected units, freed from acoustic discontinuities, can be concatenated without heavy signal processing, and their prosody, as shown by the evaluation, can be considered as natural.

Future works will be two folds. On the one hand, we will focus on the weighting to avoid the manual processing (even though it is light). On the other hand, we will focus on the building of the corpus in order to optimize the unit coverage.

”Maître corbeau, sur un arbre perché, tenait en son bec un fromage.”
/_mER@kORbo_syR9∼naRbR@pERSe_t@nEta∼so∼bEk9∼ fROmaZ_/ (french SAMPA transcription)

SAMPA	**diphone nb**				
/_mE/	2	/na/	1	/Et/	1
/Et/	1	/aR/	1	/ta s/	2
/tR/	1	/RbR/	2	/so /	1
/R@k/	2	/R@/	1	/o b/	1
/kO/	1	/@p/	1	/bEk/	2
/ORbo/	3	/pER/	2	/k9 /	1
/o_/	1	/RS/	1	/9 f/	1
/_syR/	3	/Se_/	2	/fR/	1
/R9 /	1	/_t/	1	/RO/	1
/9 n/	1	/t@nE/	3	/OmaZ_/	4

Extracts of sentences for units longer than 2 diphones :

”Mais pour lui...”
”Trente centres commerciaux...”
”Je rencontre encore beaucoup...”
”...occasion, sur les discussions...”
”...se détachent de l’arbre”
”...réseau de personnes...”
”...vous recherchez.”
”Ce poste n’existe pas non plus.”
”Attention”
”..., un bec puissant...”
”...je tiens à rendre hommage.”

Figure 4: Selected segments : lengths in diphone and extracts of sentences for units longer than 2 diphones (in the order). Written sentence is an extract of a fable of Lafontaine: The Crow and the Fox (”Master Crow perched on a tree, was holding a cheese in his beak.”)

7. References

[1] A. Black and N. Campbell, “Optimising selection of units from speech databases for concatenative synthesis,” in *Eurospeech’95*, Madrid, 1995, vol. I, pp. 581–584.

[2] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *ICASSP’96*, Atlanta, Georgia, 1996, pp. 373–376.

[3] N. Campbell and A. Black, *Prosody and Selection of source Units for Concatenative Synthesis*, pp. 279–292, New York, Springer-Verlag, 1996.

[4] M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, and S. Sandri, “Choose the best to modify the least: A new generation concatenative synthesis system,” in *Eurospeech’99*, Budapest, Hungary, 1999, pp. 2291–2294.

[5] A. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” in *Eurospeech’97*, Rhodes, Greece, 1997, pp. 601–604.

[6] A. P. Breen and P. Jackson, “Non-uniform unit selection and the similarity metric within bt’s laureate tts system,” in *ESCA/COCOSDA 3rd Workshop on Speech Synthesis*, Australia, 1998, pp. 201–206.

[7] P. Taylor and A. Black, “Speech synthesis by phonological structure matching,” in *Eurospeech’99*, Budapest, Hungary, 1999, pp. 1–25.

[8] R. Prudon and C. d’Alessandro, “A selection/concatenation tts synthesis system,” in *ISCA 4th Workshop on Speech Synthesis*, Scotland, 2001, pp. 201–206.

[9] S. Kullback and R.A. Leibler, *On information and sufficiency*, vol. 22, University of Connecticut, 1951.