A UNIT SELECTION APPROACH TO F0 MODELING AND ITS APPLICATION TO EMPHASIS

*Antoine Raux and Alan W Black**

Language Technologies Institute
Carnegie Mellon University
{antoine,awb}@cs.cmu.edu

ABSTRACT

This paper presents a new unit selection approach to F0 modeling for speech synthesis. We construct the F0 contour of an utterance by selecting portions of contours from a recorded speech database. In this approach, the elementary unit is the segment, which gives the system flexibility to combine segments from different phrases and model both macroprosody and microprosody. This method was implemented as a Festival module that can be easily reused on new voices. Using this approach, we built a model of emphasis in English. Informal experimental results show that utterances whose prosody was generated with our method are generally prefered over utterances using Festival's handwritten rule-based F0 model.

1. INTRODUCTION

1.1. Why do we need prosodic models?

Advances in concatenative speech synthesis over the past ten years have made it possible to build synthetic voices that are perfectly understandable and fairly natural [1]. One reason for the success of concatenative approaches to speech synthesis is that they circumvent the issue of prosody modeling by using portions of recorded speech *as-is*, without any prosodic modification. This results in a very natural prosody, to the extent that the system selects large enough units from its database. The price for this naturalness is a lack of control over the prosody of the generated speech. In such a framework, the prosody of a unit is tied to its phonetic content. This would not be a problem if we had a very large amount of data that covers every segment in every phonetic and prosodic context. Unfortunately, in real applications this is not the case: the database might contain some units that match very well the target utterance in terms of spectral features but not in terms of prosodic features, and vice versa. By decoupling spectral and prosodic features, we would be able to select the optimal units with regards to each aspect and use our necessarily limited resources more efficiently.

The lack of control over prosody in concatenative synthesis is even more harmful when dealing with specific prosodic cues such as the intonation patterns used to express enumerations, emphasis, or questions. These cues are very common in human speech and often crucial to proper information delivery; however, they are independent of the phonetic content of the speech. Consequently, without modification, the prosody of speech synthesized according to phonetic content is likely to be inadequate. One solution is to build synthetic voices that are specialized for each task, using a database that covers both the phonetic and prosodic patterns of the domain [2]. However, even for limited domains such as a bus schedule information system, the amount of data required to provide coverage for both phonetics and prosody is quite large. Designing and recording such a database is time, resource, and labor consuming and makes it difficult to maintain and update the system once the database has been recorded. Hence, modeling the spectral and prosodic features of speech separately would considerably reduce the amount of data required to build natural and adequate voices.

Finally, by designing databases for the sole purpose of modeling prosody, we could build prosodic models that are independent of the domain and to some extent of the speaker. When building a voice for a new task, one would only need to record a database covering the phonetic content of the domain, adequate prosody being provided by a readily available prosodic model.

1.2. Current F0 models for speech synthesis

Prosody is a combination of a number of factors such as fundamental frequency (F0), duration and pauses. In this paper, we only consider F0, which is widely recognized as the most prominent factor for the perception of prosody. The study of other aspects of prosody and of the relations among them is left as future work.

While there has been, and still is, much discussion among linguists and speech scientists on how to model F0,

*The authors would like to thank Maxine Eskenazi, Mikiko Mashimo and Brian Langner for their help with this work.

most speech synthesis systems that actually model F0 proceed in two steps: the prediction of intonational events from higher level information (e.g. semantics, syntax, discourse) and the generation of an F0 contour based on the predicted events. Intonational events are abstract labels that the system puts on syllables. They can be rather complex and include many variants (e.g. ToBI[3] uses multiple combinations of high and low tone levels) or much simpler (e.g. Tilt[4] has only two events: accent and break).

There are three types of prosodic models differing in the way they generate the F0 contour from intonation events. The first two, rule-based models and parameterized models, construct the contour according to some mathematical function. The third one, decribed in the next section, is corpus-based F0 generation, which uses natural F0 contours from databases of recorded speech.

Rule-based models use hand-written rules based on expert knowledge of prosody and the observation of some recorded data. These methods have the advantage of providing a very consistent and, if carefully designed, adequate prosody. However, manually designing a set of rules is particularly time and labor intensive so it is not easily applicable to new tasks or languages.

Another approach consists of defining parameterized curves for F0 and automatically learning the parameters from a database of recorded speech (see [5] for a description of such methods). This eliminates the need for heavy expert work when building new models and can capture more speaker-specific intonation patterns. The main problem of these first two approaches is that, although their mathematically defined F0 contours describe the general shape of the intonation, they miss a lot of fine-grained nuances that characterize natural speech. Consequently, utterances generated using these intonation models are often considered monotonous and unnatural.

1.3. Corpus-based approaches to F0 modeling

Corpus-based F0 modeling systems extract F0 contours from recorded speech databases without modifying them, so as to keep them as natural as possible. This approach is similar to concatenative segmental synthesis [6] which is known to produce more natural speech than generative synthesis. The hope is that this approach will bring the same kind of improvement to intonation modeling.

Corpus-based methods generally use databases of "templates", i.e. groups of consecutive syllables defined according to syntax (e.g. Huang et al.[7] use clauses determined by a parser) or to intonational events (e.g. Malfrere et al.[8] use intonation groups defined by the place of accents). Each template is labeled according to the sequence of intonational events marking its syllables. Given a target utterance, the system first identifies its phrases or tone groups. For each group, it finds the template whose label matches best the group's intonational events. In [7], the authors construct their template database so that only one F0 contour corresponds to each template. In [8], all the contours from the database that match a label are considered and the system selects one only when joining the contours of the different tone groups of the utterance, so as to minimize the "concatenation cost". This latter approach is to a large extent inspired by the work of Hunt and Black [6] on concatenative synthesis.

1.4. Size of the F0 selection units

By using large units (phrases), corpus-based approaches attempt to keep intact the suprasegmental structure of the utterances. By contrast, the atomic units used by concatenative segmental synthesizers are typically individual segments. The problem of using large units is that the number of such units in a reasonably sized database is restriced to at most a few thousands, which means that the F0 contour is almost uniquely predicted by the intonation events. This is a serious limitation for two reasons. First, the system might not be able to find the unit it is looking for in the database. In that case, the closest unit has to be used and eventually modified to fit the utterance. This goes against our initial claim that we want to avoid modifications of the original contours. Also, factors other than intonational events, such as syllable structure or segmental features, are known to affect F0 by producing what is called microprosody. The lack of choice among the phrasal units means that these systems will often fail to generate adequate microprosody, which could harm naturalness.

In a recent work, Meron [9] proposed a more flexible approach that uses groups of a few syllables around a single intonational event instead of larger groups. This allows his system to combine intonational events from different clauses or utterances when an exact match for the whole clause is not found.

In this paper, we propose an approach similar to Meron's but go even further in that we define the basic selection unit as a single segment. In theory this allows us to combine F0 contours from segments coming from different utterances, even inside a syllable. In practice we will see that the system almost always selects all the segments of a syllable from the same syllable in the database. This increased flexibility allows us to model F0 according to typical intonational events (in our case lexical stress, accent and pauses), syllable structure, as well as segmental features such as voicing, place and manner of articulation, etc.

2. F0 UNIT SELECTION

2.1. Unit selection in Festival

Our approach is based on the Festival speech synthesis system [10]. To a large extent, we reused its standard proce-

dures for unit selection and concatenation. For segmental unit selection, Festival labels each unit with the phoneme it represents. This defines large sets of units, each set corresponding to a given phoneme. The units of each set are then clustered according to phonetic and prosodic context [11], where context is defined by segmental features of the neighboring phonemes and suprasegmental features (e.g. stress, pauses). The clustering process is done automatically so as to minimize the acoustic distance between units of the same cluster. In practice, this creates two levels of classification of the units: one "hard", through the fixed unit labels, and one "soft", learned from the data by the clustering algorithm. We follow the same principles and define our F0 units as individual segments, classified on two levels.

2.2. F0 unit labels

Each unit is labeled by a vector whose elements are:

- word emphasis: 1 if the word containing the segment is emphasized, 0 otherwise (this feature is not used if the database does not contain emphasis labels).
- accent: 1 if the syllable containing the segment is accented as determined by Festival's intonational event prediction module, 0 otherwise.
- stress: 1 if the syllable has lexical stress as determined by the lexicon or letter-to-sound rules, 0 otherwise.
- syllable position: `single` if the word containing the segment is monosyllabic, `initial` if the syllable containing the segment is word-initial, `medial` if it is word-medial, `final` if it is word-final.
- nature of the following syllable break: 0 if the syllable is not followed by a word boundary, 1 if it is followed by a word boundary, 2 if it is followed by a phrase boundary and 3 if it is followed by a sentence boundary.
- syllable structure: `V` if the syllable containing the segment is a single vowel, `CV` if it is a vowel preceded by consonants, `VC` if it is a vowel followed by consonants, and `CVC` if it is a vowel both preceded and followed by consonants.
- position in syllable: `onset` if the segment is in the onset of a syllable, `coda` if it is in the coda of a syllable.

This choice of features is partly based on the work of Imoto et al. [12] on the automatic recognition of syllable stress level in spoken English. They established that training separate models according to syllable structure and syllable breaks significantly improved stress classification accuracy. We conclude that different syllable structures and breaks yield different microprosody. Thus, we explicitly integrate these features in the unit labels, which represent the "hard" classification of the units.

In addition, note that although this is not strictly enforced, the constraints imposed by the last two features, along with the weight of the concatenation cost (see section 2.4) strongly bias the system towards selecting full syllables from the database (i.e. all segments from the same syllable).

2.3. F0 unit clustering

As for segmental unit selection, we perform clustering over the sets of units bearing the same label. The clustering algorithm requires two sets of features: the features on which the clustering decisions are made (the "context") and a measure of distance between two units. Since Festival's clustering algorithm is able to automatically select the features that are most useful, we initially provide an extensive set of features containing:

- segmental features of the target segment (phoneme name, voicing, place and manner of articulation).
- segmental features of the neighboring segments.
- nature of the four neighboring syllable breaks (two before, two after).
- stress of the four neighboring syllables.
- accent of the four neighboring syllables.
- estimated part-of-speech of the word containing the segment and the neighboring words.

To measure the acoustic distance between two units, we extract the F0 and ΔF0 values every 5 ms. This defines a set of 2-dimensional vectors whose size depends on the length of the unit. The distance between two units is then computed using the Mahalanobis metric described in [11]. This is the same method that Festival uses to compute the distance between units when doing concatenative segmental synthesis, except that we use F0 and ΔF0 instead of cepstral coefficients, power and their deltas.

2.4. Synthesis using the F0 model

Given an input sentence, eventually augmented with meta-information such as emphasis, Festival first performs text analysis and extract segment-, syllable-, word-, and phrase-level features. Based on these features, the system determines the F0 unit label corresponding to each segment of the utterance, and identifies the cluster matching the segment's context. From that cluster, Festival selects the unit that minimizes an overall cost function combining a target

cost (how far is the unit from the center of its cluster?) and a concatenation cost (how well does this unit join with the previous one?). This is again the same method that is used for segmental synthesis (see [11] for details), except that, here, the costs are only dependent on F0 and ΔF0 instead of cepstral parameters, power and their deltas.

Once the utterance's F0 contour is built, it is applied to the synthesized waveform through LPC modification. The waveform can be generated either from the same data as the F0 contour or using a different unit selection or diphone voice. However, we currently don't perform any normalization of the F0 values so the target segmental voice must have a pitch range similar to that of the voice used for F0 modeling. In the future, we plan to normalize all values according to F0's mean value and standard deviation (z-scores), which will allow easy transfer of F0 models from one speaker to another.

2.5. Implementation issues and integration in Festival

One problem with the method described above is that the units selected for F0 do not necessarily have the same duration as the corresponding selected segments. To solve this issue, we modify the time stamps of the F0 values extracted from the selected units linearly. Thus, the extracted portions of F0 curves are stretched or contracted to fit the duration of the segments.

We also tested the impact of smoothing on our model. To do so, instead of applying the F0 contour as-is, we select some points (one every 40ms) and take them as target points between which we let Festival linearly interpolate the F0 curve. As can be seen in Figure 1, there is a clearly visible difference between smoothed and non-smoothed contours. However, this difference is hardly perceptible to the ear because discontinuities almost always occur at syllable boundaries. This confirms that our method tends to select whole syllables from the database.

In order to test our approach and make it easy to apply to new voices, we implemented it as a set of scripts that build and run "F0 voices" in Festival. An F0 voice is built using a script similar to that used to build segmental unit selection voices. The F0 voice is then accessible as a standard F0 model. We also provide the capacity to use the F0 contours on one of Festival's default voices. In the future, our goal is to package F0 voices as distinct models that can be imported in any unit selection or diphone voice.

3. APPLICATION TO GENERAL F0 MODELING

3.1. The CMU Arctic Database

In order to build a general F0 model for English, we applied our approach to the CMU Arctic database [13], a new, freely available database of recorded speech designed for

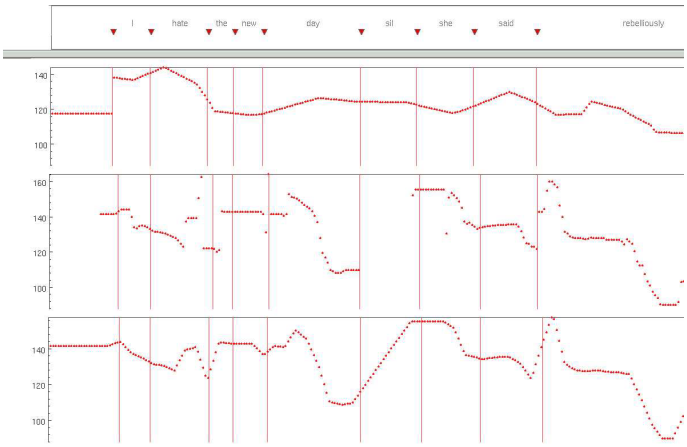

Fig. 1. The F0 contours generated by hand-written rules (top), F0 unit selection (middle) and F0 unit selection with smoothing (bottom), for the sentence "I hate the new day, she said rebelliously". Vertical lines are word boundaries.

unit selection speech synthesis research. This database was designed and recorded at Carnegie Mellon University, and is distributed through the Festvox website. Specifically, we used the recordings of the male Scottish speaker (awb). The database consists of around 1200 utterances designed to offer a good phonetic coverage. The total number of units, including pauses, is about 41000. The recordings were automatically labeled using the CMU Sphinx speech recognizer using the Festvox scripts. No hand correction of the labels has been made.

3.2. Evaluation

We trained a CART tree duration model on the database, which is a standard procedure in Festival. We also built a segmental unit selection voice on the database, along with our F0 voice. We then compared utterances generated using the duration model, segmental and F0 voice with the same utterances generated with the same duration model and segmental voice but using Festival's standard rule-based F0 model.

Although there are still some cases where the F0 unit selection produces inadequate prosody, in most cases it is better than the rule-based model. In particular, when listening to long series of sentences, the rule-based model tends to produce very monotonous intonation patterns. The prosody of each sentence sounds similar to the others. By contrast, F0 unit selection produces more varied pitch contours, depending on the prosodic and phonetic context. It also makes use of a wider range of F0 values, as can be seen on the example in Figure 1. Smoothing did not affect the results significantly.

We conducted an informal blind test where 4 subjects

Model	General	Emphasis
F0 Unit Selection	11	10
Rule-based F0	2	0
Neither	12	4

Table 1. Comparison of F0 unit selection with a rule-based F0 model. The figures are the number of sentences for which the model obtained at least 3 votes out of 4. Sentences where no model obtained more than 2 votes are counted as "neither".

listened to 25 sentences, each in two versions, one using our smoothed F0 model and one using the rule-based model. They were then asked to say which version they prefered, or neither if they did not have any preference. For each sentence, we counted the number of votes for each model. The results, given in Table 1, indicate that prosody generated by F0 unit selection was prefered in almost half of the sentences. By contrast, rule-based prosody got a majority of the votes in only 2 cases. Hence, we conclude that our model performed at least as well as, and often better than, the rule-based model.

4. APPLICATION TO EMPHASIS MODELING

4.1. Database of emphasized speech

To test our approach on a specific prosodic phenomenon, we built a model of F0 for sentences containing emphasized words. We used a database specifically designed to study emphasized speech, provided by Cepstral LLC [14], a Pittsburgh-based company specialized in building synthetic voices. The data consists of 547 English sentences read by the same Scottish speaker as the CMU Arctic database described in section 3. 270 sentences are read naturally. For the remaining 277 sentences, the speaker emphasized every other word in the sentence. Although each word is emphasized in a natural way, the abundance of emphasized words results in sentences that are somewhat unnatural and hard to understand. However, the advantage of this approach is that it provides a large number of emphasized words in a relatively small number of sentences. The total number of emphasized words is 968, with 505 unique words. These words cover a wide range of word length (from monosyllabic words such as "if", "you" or "fault", to 5-syllable words such as "philosophical"), as well as all the most common syllable structures of English. In total, the database contains approximately 16000 units (including pauses). The recordings were automatically labeled using the CMU Sphinx speech recognizer without hand correction.

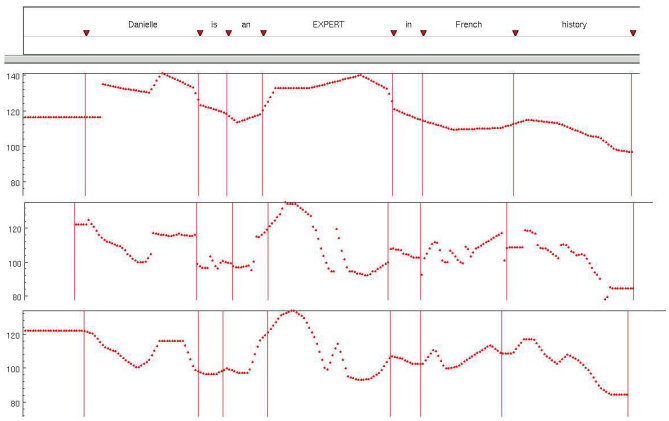

Fig. 2. The F0 contours generated by hand-written rules (top), F0 unit selection (middle) and F0 unit selection with smoothing (bottom), for the sentence "Daniele is an expert in French history." with emphasis on "expert".

4.2. Evaluation

We compared the utterances generated using the resulting F0 model with the same utterances using Festival's standard rule-based model of emphasis. In both cases, the underlying voice was a diphone voice built using a different speaker than the one used for the emphasis database (but with a similar pitch range). In general, our model gave much more natural prosody than the rule-based model. In particular, it was able to produce natural emphasis independently of the position of the emphasized word in the sentence (initial, medial or final). This shows that the model (in particular the clustering algorithm) was able to capture the differences in pitch curve associated with different word positions. Also, the model worked well for words having various lexical stress patterns. Again, we explain this by the fact that the model was able to characterize the pitch curves of a wide variety of emphasized words found in the database. Figure 2 shows the contour generated by the rule-based model and by our model with and without smoothing, on an utterance containing an emphasized word. As for general prosody, it appears that our approach produces a wider dynamic range than the rule-based method, while still keeping prosody natural. Again, smoothing did not seem to have an effect on auditory perception on our test sentences.

We confirmed our impressions by performing an informal blind test on the same 4 subjects who evaluated the general F0 model. They listened to 14 sentences, each in two versions, one whose pitch contour was generated by the rule-based model and one by our method (smoothed). The utterances were constructed to contain words that could be naturally emphasized. The results are shown in Table 1. For 10 sentences, at least 3 subjects prefered the prosody generated by F0 unit selection. For the 4 remaining sentences, no

model got a majority of the votes, and there was no sentence where the rule-based prosody got a majority of the votes.

The main limitation of our emphasis model comes from the way the database was designed. Since sentences that contain emphasized words are "artificially" emphasized every other word, it is not possible to model the natural prosody of non-emphasized words in a sentence containing an emphasized word. Another issue is that we only model emphasis when it affects single words. More data would be needed to model F0 on compound words or phrases. By designing a database that contains naturally emphasized sentences, we should be able to capture finer nuances and produce even better contours.

From this evaluation, it appears that our method provides a very efficient way to build natural F0 models for specific aspects of prosody. In the future, we plan to try it on other phenomena such as prominence or on different speaking styles. In each case, it only requires to design and record a database for our specific needs, along with some minor changes in the model (such as adding a feature to characterize the degree of prominence of a word).

5. CONCLUSION

In this paper, we presented a new approach to F0 modeling based on the concatenation of F0 contours from a database of recorded speech. By using individual segments as units, our approach provides maximal flexibility in unit selection and takes into account a wide range of features at the phrase, word, syllable and segment levels. We believe that this flexibility gives the model the ability to render both macroprosodic and microprosodic events, resulting in increased naturalness. Being fully data-driven, this method offers a cost-effective way to build natural F0 models, both for general purposes and for specific domains, speakers, speaking styles or prosodic phenomena.

6. ACKNOWLEDGMENTS

This material is based upon work supported in part by the U.S. National Science Foundation under Grant No. 0208835, "LET'S GO: improved speech interfaces for the general public". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] A. Black, "Perfect synthesis for all of the people all of the time," in *IEEE TTS Workshop*, Santa Monica, CA, 2002.

[2] A. Black and K. Lenzo, "Limited domain synthesis," in *ICSLP2000*, Beijing, China., 2000, vol. II, pp. 411–414.

[3] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *ICSLP '92*, Banff, Canada, 1992, pp. 867–870.

[4] P. Taylor, "The tilt intonation model," in *ICSLP '98*, Sydney, Australia, 1998.

[5] A. Syrdal, G. Moehler, K. Dusterhoff, A. Conkie, and A. Black, "Three methods of intonation modeling," in *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 305–310.

[6] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP '96*, Philadelphia, PA, 1996, pp. 373–376.

[7] X. Huang, A. Acero, J. Adcock, H. Hon, J. Goldsmith, J. Liu, and M. Plumpe, "Whistler: A trainable text-to-speech system," in *ICSLP '96*, Philadelphia, PA, 1996.

[8] F. Malfrere, T. Dutoit, and P. Mertens, "Automatic prosody generation using supra-segmental unit selection," in *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 323–328.

[9] J. Meron, "Prosodic unit selection using an imitation speech database," in *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland, 2001.

[10] *The Festival Speech Synthesis System*, www.cstr.ed.ac.uk/projects/festival/.

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

[12] I. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system," in *ICSLP '02*, Denver, CO, 2002, pp. 749–752.

[13] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[14] *Cepstral LLC*, www.cepstral.com.