Modelling F_0 Dynamics in Unit Selection Based Speech Synthesis*

Daniel Tihelka, Jindřich Matoušek, and Zdeněk Hanzlíček

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitni 8, 306 14 Plzeň, Czech Republic
{dtihelka,jmatouse,zhanzlic}@kky.zcu.cz

Abstract. In the common unit selection implementations, F_0 continuity is measured as one of concatenation cost features with the expectation that smooth units transition (regarding speech melody) is ensured when the difference of F_0 is low enough. This measure generally uses a static F_0 value computed at the units boundary. In the present paper we show, however, that the use of static F_0 values is not enough for smooth speech units concatenation, and that a dynamic nature of the F_0 contour must be taken into account. Two schemes of dynamic F_0 handling are presented, and speech generated by both schemes is compared by means of listening tests on specially selected phrases which are known to carry unnatural artefacts. Advantages and disadvantages of the individual schemes are also discussed.

Keywords: text-to-speech synthesis, unit selection, concatenation cost, fundamental frequency F_0.

1 Introduction

There have been many papers describing concatenation cost features in unit selection speech synthesis, [20,21,15,4,16,1,13,14] to name a few. While most of them aim at determining the sources of spectral discontinuities, with results often in contradiction, in [5] was shown that a large number of audible discontinuities tend to appear at joins with incoherent F_0 values in the wider area around concatenation points.

There is a general agreement across unit selection researches that the incorporation of F_0 continuity measure at the units boundaries is an essential condition of smooth concatenation achievement. Usually, however, the authors limit this feature to simple "static" F_0 difference in Hz (or log(Hz) in some cases) [3,2,12], i.e. $d = \left|f^e(i) - f^b(i+1)\right|$, where $f^e(i)$ denotes the F_0 value assigned to the end of the i^{th} unit, and $f^b(i+1)$ value assigned to the beginning of the $(i+1)^{\text{th}}$ unit. The manner of F_0 computation may differ (and it usually does) for individual approaches, but it basically is an average through several epoch periods to eliminate the F_0 fast changes

* The research has been supported by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090, and by the Technology Agency of the Czech Republic, project No. TA01011264.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 457–464, 2014.

in microprosody. Nevertheless, whatever the F_0 computing scheme, it must be ensured that $f^e(i) = f^b(i+1)$, and thus $d = 0$, for the two units following each other in the speech corpus.

In this paper, the preliminary experiments taking into account wider F_0 context are described and discussed. In our work we extend [8], but instead of evaluating specially designed phrases with a single concatenation point in the middle of vowels [5], we will employ a real TTS system on which the results are obtained.

2 The Ways of F_0 Dynamics Modelling

First, let us describe what the baseline implementation of concatenation cost computation looks like in our TTS system ARTIC [11,6]. When concatenating two units (diphones in our case) i and $i+1$, the concatenation cost $C^c(i, i+1)$ is computed as

$$C^c(i, i+1) = \frac{C^c_S(i, i+1) + C^c_E(i, i+1) + C^c_F(i, i+1)}{3} \tag{1}$$

where $C^c_S(i, i+1)$ is the Euclidean distance of 12 MFCC coefficients expecting to reflect spectral smoothness of the concatenated units[1], $C^c_E(i, i+1)$ is the absolute difference of energy, and $C^c_F(i, i+1)$ reflects the "static" F_0 difference at the units boundaries, computed according to Equation (7) as $C^c_F(i, i+1) = \left|\delta\big(f^e(i), f^b(i+1)\big)\right|$.

All the features are computed in the pitch-synchronous way, meaning that each pitch-mark (see [7] for the definition) has been assigned the value of energy, F_0 (being `NaN` for unvoiced pitch-marks), and the vector of MFCC coefficients. All the values are z-score normalized to align their ranges. Then, each unit boundary obtained by HTK-alignment process [10,9] is tied with the set of features computed for the pitch-mark being the closest to the given boundary.

While both energy and MFCC are computed from a window of fixed length, centred around the pitch-mark the resulting value is assigned to, the computation of F_0 is slightly more complicated. For a sequence of voiced pitch-marks $p(k), k = 1, 2, \ldots, K$, each pitch-mark has assigned mean F_0 value $f(k)$:

$$f(k) = \frac{\sum\limits_{l=x}^{y-1} \frac{1}{p(l+1)-p(l)}}{y - x} \tag{2}$$

$$x = k - w\lfloor \frac{k}{K} + 0.5 \rfloor \tag{3}$$

$$y = k + w\lfloor \frac{K-k}{K} + 0.5 \rfloor \tag{4}$$

where w is the fixed number of epochs through which the F_0 is computed. As illustrated on Figure 1, it enables the use of a fixed number of epochs regardless of whether the F_0 value is computed at the beginning, middle, or at the end of the voiced pitch-marks sequence.

[1] The use of MFCC is revised currently, since our evidence suggests that it does not seem to be an appropriate feature for such measure. Therefore, although it is used in this experiment, it may become invalid in foreseeable future.

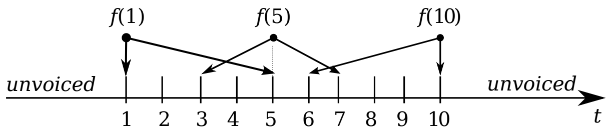

Fig. 1. The illustration of pitch-synchronous F_0 values computation. $K = 10$, $w = 4$, vertical lines represent pitch-marks.

In this paper we experiment only with values of $C^c_F(i, i + 1)$ in Equation (1). The remaining features, as well as the manner of target cost computation, stay untouched.

2.1 Delta Coefficients

Natural consideration of F_0 dynamics employment is to use "classic" *delta* coefficients. However, they are not very suitable for unit selection synthesis, since they reflect the dynamics only in a relatively near point around the concatenation point. Although such dynamics are usually used together with spectral and other features in HMM synthesis (where their use is legitimate since HMM works as generator on model states), their use for longer cross-unit F_0 fluency policing is not very effective.

To illustrate it, let us take HTK [22] toolkit as an example. For k^{th} feature value, its dynamic coefficients are computed as:

$$D(k) = \frac{\sum_{i=1}^{I} i \Big(F(k+i) - F(k-i) \Big)}{2 \sum_{i=1}^{I} i^2} \tag{5}$$

where I is the configurable length of window through the dynamics are computed and F is the value of the feature (F_0 in out case). It is obvious that even with $I > 1$, the largest portion of the delta value is taken from the difference to $k - 1$ and $k + 1$ point. And the same situation is for the *acceleration* (*delta-delta*) coefficients, which are computed by Equation (5), except using the computed D in place of F.

Contrary to this, we aimed at involving the wider tendency of F_0 behaviour, since it is natural supra-segmental feature expressing the communication function of a phrase crossing several adjacent phones. On the other hand, the considered context must not be too wide (e.g. crossing several diphones) since the feature would not reflect what it is intended to (i.e. the smoothness of join), but it would instead describe something like supra-segmental prosody tendencies, while a local audible unnatural artefact could still happen.

2.2 Contour Comparison

Quite encouraging results were reported in [8], in which the vector of 8 F_0 values extracted from the vicinity of carefully designed concatenation point is able to detect audible discontinuities with accuracy about 90%. Therefore, we were curious whether the scheme is able to provide a similar result when employed in the real TTS system.

To compare the contour F_0 values, each unit boundary was extended with 9 F_0 values. That is, $f^b(i,k), k = 1.2, \ldots, 9$ values were assigned to the beginning of i^{th} unit with $f^b(i,5)$ equal to the beginning of the unit, and similarly, $f^e(i,k), k = 1.2, \ldots, 9$ are assigned to the end of unit equal to $f^e(i,5)$; see Figure 2 for the illustration. Thus, the requirement $f^e(i,k) = f^b(i+1,k)$ from Section 1 is still valid $\forall i, k$ for adjacent units, while the context 9 pitch-marks long is compared in the concatenation cost.

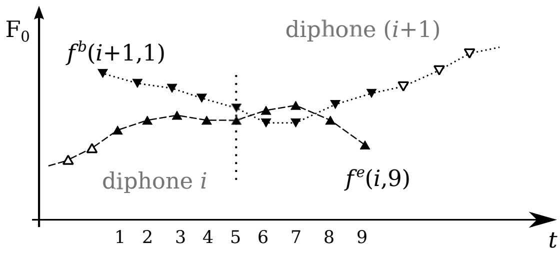

Fig. 2. The illustration of F_0 contour computation on a join of voiced diphones boundary. The dashed line connects values $f^e(i)$, dotted connects values $f^b(i+1)$, black triangles represent the $k = 1.2, \ldots, 9$ pairs of the F_0 values used in Equation (6). Concatenation point is dotted vertical line.

The value of F_0-related sub-cost is computed here using Euclidean distance between the corresponding f values as:

$$C_F^c(i, i+1) = \sqrt{\sum_{k=1}^{9} \delta\Big(f^e(i,k), f^b(i+1,k)\Big)^2} \tag{6}$$

where $\delta(a, b)$ is function defined as:

$$\delta(a,b) = \begin{cases} a - b, & a \neq \texttt{NaN},\ b \neq \texttt{NaN} \\ 0, & a = b = \texttt{NaN} \\ 6, & \text{otherwise} \end{cases} \tag{7}$$

with the value 6 chosen as large enough, since the difference of z-score normalized values f will exceed it for large F_0 differences only (exactly it is in case $a \leq -3,\ b \geq 3$, each having 0.1% likelihood). However, the particular value does not matter a great deal.

2.3 Slope

The main disadvantage of F_0 contour comparison scheme is its higher computation cost — there are 9 floating points multiplications followed by square root evaluation. Considering the number of evaluations which are carried out during the concatenation cost computing (may approach 250 millions, as described in [18]), such a scheme will have a significant negative impact on the performance of the TTS system, which would be notable especially on lower-resource devices, e.g. smart-phones [19].

This is the reason why we have experimented with another scheme of F_0 dynamics embedding – the comparison of the F_0 slope of the concatenated units; in [8] it was reported as only slightly worse than the use of contour. Firstly, we have computed the slope of F_0 using the linear regression of all the F_0 values measured on a voiced *phone* (such covered by voiced pitch-marks in more than 70% of length[2]); let us mark it as $S(j)$, where j is the index of phone within a phrase. Then, the sequence of phones is converted to *diphones*, so two values $S(j)$, $S(j+1)$ are assigned to diphones as $S(j) = S^e(i-1) = S^b(i)$, $S(j+1) = S^e(i) = S^b(i+1)$; the whole scheme is illustrated on Figure 3. The value of F_0-related sub-cost is then computed as:

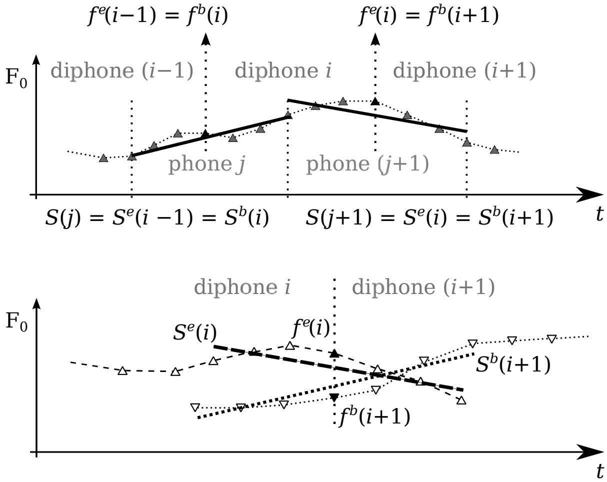

Fig. 3. The illustration of F_0 slope (line) computation and its phone-to-diphone distribution in the upper part of the Figure. In the lower part, the dashed line represents F_0 contour ($f^e(i) = \blacktriangle$) and slope $S^e(i)$, dotted line illustrates contour ($f^b(i+1) = \blacktriangledown$) and slope $S^b(i+1)$ used in Equation (8). Concatenation point is dotted vertical line.

$$C_F^c(i, i+1) = \begin{cases} 3\left|f^e(i) - f^b(i+1)\right| + 2\left|S^e(i) - S^b(i+1)\right|, & f^e(i) \neq \mathtt{NaN}, \\ & f^b(i+1) \neq \mathtt{NaN} \\ 0, & f^e(i) = \\ & f^b(i+1) = \mathtt{NaN} \\ 15 = 9 + 6, & \text{otherwise} \end{cases} \tag{8}$$

[2] This is slight difference from [8] when the slope was computed only through 4 pitch-periods around the concatenation point

where the first part is the difference of z-score F_0 values at the diphone boundaries computed exactly as in the baseline system, and the second part is the difference of F_0 slopes. The weights were chosen to slightly prefer the static F_0 value to the slope.

3 Evaluation

To evaluate the effect of F_0 dynamics modelling, we have designed listening tests with 20 phrases taken from the set in which the largest number of unnatural artefacts were evaluated in our internal research. There were 14 people involved in the test for which 3-point scale CCR (comparison category rating) form was used. The pairs compared were *baseline* × *slope* and *contour* × *slope*, presented in the randomized ordering. The *baseline* × *contour* test had been carried out earlier during the research to clarify results from [8], but lower number of listeners participated in it. Therefore, although we present the results of this test as well, and the tendency they display is in agreement with the overall results, note that they are not fully comparable with the main tests.

Table 1. The comparison of preference of the individual system versions

Test (A × B)	Prefer A	No preference	Prefer B
contour × *slope*	38.8%	36.9%	24.3%
baseline × *slope*	16.5%	43.5%	40.1%
baseline × *contour*	9.0%	34.0%	57.0%

It can clearly be seen that while *contour*-incorporating version is generally preferred, both versions are preferred to the baseline system, where only the difference of unit boundaries-related "static" F_0 is computed. When comparing *contour* to *slope*, there is slight preference for the use of *contour*; we expect that the reason is more precise F_0 contour comparison. On the other hand, this computation scheme is much more demanding, as mentioned in Section 2.3. It may seem that the use *contour*-based dynamics are evidently more preferred to the *baseline* that the *slope* is, but note again that this test has not been carried out by the same number of listeners, although on the same set of phrases.

4 Conclusion

The results presented are in general agreement with the results of [8], so it may be concluded that the use of dynamic F_0 features as a part of the concatenation cost has noticeable effect on the quality of speech synthesis. What remains to be found is the most effective, both in terms quality improvement and computation speed, scheme of the features comparison. We plan experiments, where, for example, fewer F_0 points will be compared in the *contour* scheme, or where the Euclidean distance in Equation (6) will be replaced by the mean of $\left|\delta\big(f^e(i,k),\, f^b(i,k)\big)\right|$ absolute differences. We also need to check the slope computed exactly as described in [8].

Moreover, as a part of listening tests stimuli, we plan to use phrases where clear F_0 artefact is found, when generated by baseline system. Currently, although the evaluated phrases do contain unnatural artefacts, they may be of any type. Due to the rather small range of listening test and the fact that only phrases used for evaluation have been synthesized, we did not also carry out the evaluation of results reliability, as described in [17]. We plan to do so in the near future.

Special thanks are due to National Grid Infrastructure MetaCentrum, providing the access to computing and storage facilities under the program LM2010005 "Projects of Large Infrastructure for Research, Development, and Innovations".

References

1. Bellegarda, J.R.: A novel discontinuity metric for unit selection text-to-speech synthesis. In: Proc. of 5th Speech Synthesis Workshop (SSW5), Pittsburgh, PA, USA, pp. 133–138 (2004)
2. Conkie, A., Syrdal, A.K.: Using F0 to constrain the unit selection Viterbi network. In: Proc. of Acoustics, Speech, and Signal Processing ICASSP, pp. 5376–5379. IEEE (2011)
3. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. of Acoustics, Speech, and Signal Processing ICASSP 1996, vol. 1, pp. 373–376. IEEE (1996)
4. Klabbers, E., Veldhuis, R.N.J.: Reducing audible spectral discontinuities. IEEE Transactions on Speech and Audio Processing 9(1), 39–51 (2001), http://dblp.uni-trier.de/db/journals/taslp/taslp9.html#KlabbersV01
5. Legát, M., Matoušek, J.: Design of the test stimuli for the evaluation of concatenation cost functions. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 339–346. Springer, Heidelberg (2009)
6. Legát, M., Matoušek, J.: Collection and analysis of data for evaluation of concatenation cost functions. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 345–352. Springer, Heidelberg (2010)
7. Legát, M., Matoušek, J., Tihelka, D.: On the detection of pitch marks using a robust multiphase algorithm. Speech Communication, 552–566 (2011), http://www.kky.zcu.cz/en/publications/LegatM_2011_Onthedetectionof
8. Legát, M., Matoušek, J.: Pitch contours as predictors of audible concatenation artifacts. In: Proc. of World Congress on Engineering and Computer Science 2011, San Francisco, USA, pp. 525–529 (2011)
9. Matoušek, J., Romportl, J.: Automatic pitch-synchronous phonetic segmentation. In: INTERSPEECH 2008, Proc. of 9th Annual Conference of International Speech Communication Association, Brisbane, Australia, pp. 1626–1629 (2008)
10. Matoušek, J., Tihelka, D., Psutka, J.V.: Experiments with automatic segmentation for Czech speech synthesis. In: Matoušek, V., Mautner, P. (eds.) TSD 2003. LNCS (LNAI), vol. 2807, pp. 287–294. Springer, Heidelberg (2003), http://dx.doi.org/10.1007/978-3-540-39398-6_41
11. Matoušek, J., Tihelka, D., Romportl, J.: Current state of Czech text-to-speech system ARTIC. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 439–446. Springer, Heidelberg (2006), http://dx.doi.org/10.1007/11846406_55
12. Narendra, N.P., Rao, K.S.: Syllable specific unit selection cost functions for text-to-speech synthesis. ACM Transactions on Speech and Language Processing 9(3), 5:1–5:24 (2012), http://doi.acm.org/10.1145/2382434.2382435

13. Pantazis, Y., Stylianou, Y.: On the detection of discontinuities in concatenative speech synthesis. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) COST 277. LNCS, vol. 4391, pp. 89–100. Springer, Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-71505-4_6
14. Přibil, J., Přibilová, A.: Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech. EURASIP Journal on Audio, Speech, and Music Processing 33(3), 1–22 (2013), http://dx.doi.org/10.1186/1687-4722-2013-8
15. Stylianou, Y., Syrdal, A.K.: Perceptual and objective detection of discontinuities in concatenative speech synthesis. In: Proc. IEEE Acoustics, Speech, and Signal Processing (ICASSP), pp. 837–840 (2001)
16. Syrdal, A.K., Conkie, A.D.: Data-driven perceptually based join costs. In: Proc. of 5th Speech Synthesis Workshop (SSW5), Pittsburgh, PA, USA, pp. 49–54 (2004)
17. Tihelka, D., Grůber, M., Hanzlíček, Z.: Robust methodology for TTS enhancement evaluation. In: Habernal, I. (ed.) TSD 2013. LNCS, vol. 8082, pp. 442–449. Springer, Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-40585-3_56
18. Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi search for fast unit selection synthesis. In: INTERSPEECH 2010, Proc. of 11th Annual Conference of the International Speech Communication Association, pp. 174–177 (2010), http://www.isca-speech.org/archive/interspeech_2010/i10_0174.html
19. Tihelka, D., Stanislav, P.: ARTIC for assistive technologies: Transformation to resource-limited hardware. In: Proc. of World Congress on Engineering and Computer Science 2011, San Francisco, USA, pp. 581–584 (2011)
20. Vepa, J., King, S.: Kalman–filter based join cost for unit–selection speech synthesis. In: Proc. EUROSPEECH 2003 – INTERSPEECH 2003, Proc. of 8th European Conference on Speech Communication and Technology, pp. 293–296. ISCA (2003)
21. Vepa, J., King, S.: Join cost for unit selection speech synthesis. Ph.D. thesis, The University of Edinburgh, College of Science and Engineering, School of Informatics (2004), https://www.era.lib.ed.ac.uk/handle/1842/1452
22. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.4. Cambridge University Press (2006)