F0 contour and segmental duration modeling using prosodic features

Agnieszka Wagner, Katarzyna Klessa

Department of Phonetics, Adam Mickiewicz University in Poznań, Poland
{wagner, klessa}@amu.edu.pl

Abstract

This paper proposes a framework of F0 contour generation and segmental duration modeling for application in a unit-selection speech synthesis system for Polish – BOSS. We describe the design of the F0 and duration modeling modules and emphasize the role of prosodic features (related to stress, pitch accent and phrase) in these two tasks.

Index Terms: intonation modeling, F0, duration, prosodic features, speech synthesis

1. Introduction

For an effective use in speech technology applications the framework of F0 and duration modeling should provide means of generating *high-quality, comprehensible* and *natural* F0 and timing patterns from a linguistically meaningful representation (i.e. containing information significant to the linguistic interpretation) of an utterance's prosody. The representation should be automatically derivable from utterance's acoustics and annotation. Apart from that the prosody modeling framework should be efficient, which entails *ease of implementation, low computational load* and *automated data-driven training* without the need of extensive hand-tuning. To achieve this objective it is necessary to build a framework which relies on *a small number of models* trained to perform a *reliable* and *accurate* estimation of F0 contour representation from which the contour can be generated in few steps.

The state of the art models are well capable of generating a naturally sounding intonation and timing patterns in neutral (i.e. news commentary style) speech, but duration and F0 modeling in expressive and emotional speech still remains a challenge. Therefore, one of the assumptions underlying the framework proposed in this study is *independence from speech style.* To achieve this goal it appears necessary to investigate the relationships among the low level phonetic-acoustic features as well as suprasegmental prosodic features including the possible interactions between various levels of analysis under specific conditions of various speech styles. The important factor is the applicability of the framework to F0 and duration modeling in neutral and expressive speech. There are two major approaches to F0 contour generation: *sequential* and *superpositional*. The former assumes that F0 contours can be modeled as *sequences* or *strings* of intonational events (such as pitch accents, boundary tones) associated with specific structures (accented syllable) or landmarks (e.g. phrase edge) in the segmental string [1, 2, 3]. On the contrary, in superpositional models F0 contours are regarded as a result of *superposition* of two or more components of a different temporal scope e.g. [4]. Most of the sequence-based models rely on interpolation either between pitch targets anchored in syllable structure [5, 6] or derived from parameters describing intonational events on the phonetic level [2, 3, 7, 8].

Duration prediction models for speech technology range from the more traditional, rule-based techniques [9] to trainable, corpus-based techniques. Nowadays, it is often the case that the two approaches overlap and careful linguistic feature extraction usually is an important stage preceding the actual statistical processing [10, 11, 12]. The unit that should be regarded as the base for segmental duration modeling is also a subject of discussion. Most frequently, phone is used as the unit, however there are also alternative proposals e.g. Campbell's syllable-based model [13]. Linguistic knowledge may be used not only in the data preparation process but also in the modeling process itself which is postulated and tested for various languages by Van Santen's sum-of-product models [14, 15]. As for the Polish segmental duration, certain important statements were already made in quite early publications [16, 17], however previous research on segmental duration for Polish was based mostly on smaller amounts of data containing mostly isolated words or phrases.

1.1. Features of the present framework

1.1.1. F0

The framework proposed in the current study will follow the assumptions underlying the sequential approach, but some reference to superpositional features will be made as well. We will follow the target-based approach (interpolation between pitch targets of a constant position in syllable) rather than the parameter-based one, because it involves a smaller number of intermediate steps between pitch target estimation and contour generation and thus, can be regarded as more efficient.

However, contrary to other authors [5, 6] we train one regression model instead of three models to estimate the level of three pitch targets: at the start (*f0start)* and end of the syllable (*f0end)*, and in the middle of the nucleus (*f0mid*). One regression model is designed for neutral speech and the other for expressive speech. The models estimate relative F0 values (z-score normalized with respect to mean F0 level determined for a given phrase type: initial, medial, final or single, cf. [18]) and rely on a linguistically meaningful representation of utterance's intonation in which prosodic features play a primary role. F0 contour generation is carried out in three steps: a) re-scaling of the estimated pitch targets, b) determination of the temporal alignment of the targets and c) smoothing and interpolation between the targets.

1.1.2. Duration

It was assumed that segmental duration may be simultaneously modified by a set of features from both segmental and suprasegmental levels. The initial selection of features was based on a series of statistical significance tests and after that the correlation of the obtained set was calculated using Classification and Regression Trees method (CART). Individual phones were the base units for prediction but the influence of other units was not neglected, actually - among others, the choice of the features for duration prediction was motivated by the intention of finding the relationships that might indicate another unit which might be more suitable to serve as a base for prediction.

With the present data it became possible to observe the

relations between segmental duration and factors both from the segmental and suprasegmental viewpoint, since the speech corpus delivers a variety of texts ranging from short phrases, through longer and more complex sentences up to continuous text including both formal and informal speech styles.

2. Methodology

1.2. Speech material

The substantial part of the speech corpus used in the current study was built for the Polish module of BOSS (Bonn Open Source Synthesis) which is a unit selection speech synthesis system. The corpus contains recordings of a professional male speaker (approx. 2 hours) reading phonetically rich and balanced sentences, fragments of fiction (including dialogues and examples of expressive speech) and reportage.

Additionally, for F0 modeling in expressive speech a speech corpus containing recordings of three speakers (two female and one male) reading a novel chapter was used. The text included examples of different discourse types, dialogues, monologues, modes and expressivity. The speakers gave it a very lively and emotional interpretation; 20 minutes of speech per speaker was obtained (1758 phrases, 9722 syllables).

For duration prediction apart from the BOSS speech material, a reference set of 40 voices reading a short paragraph of continuous speech was used to verify the results from BOSS prediction with data from a larger group of speakers. The 25-sentence paragraph delivers descriptive phrases as well as examples of dialogue covering the most common Polish phrase types. The speakers were twenty man and twenty women approximately of the same age (20-25 years).

1.3. Labeling procedure and feature extraction

Transcription and segmentation at the phone, syllable and word level was obtained automatically using SALIAN software [19]. At the word level, verification of the automatically inserted stress markers was carried out with the help of a large pronunciation lexicon and five pitch accent and seven boundary tone types were marked manually. On the basis of the information provided in the labeling a number of syllable- and phrase-level features was extracted automatically from the speech signals with a *Praat* script. The features which had statistically significant effect on duration and level of f0start, f0mid and f0end targets were used as predictor variables to train statistical models.

The duration extraction from the annotated material was performed with a specifically designed software. Then the extracted and tagged data was exported to *Statistica*. The first step of the duration analysis focused on the distributions, means, and variances of the duration as a variable dependent on a presumed set of modifying factors. In the second step, the usefulness of a set of 57 modifying factors for duration prediction was assessed by means of the CART algorithm [19]. The results supported the claim that the duration of speech sounds may be modified by the influence of segmental and suprasegmental features as well as by their combination.

1.4. Linguistic representation

1.4.1. F0 target estimation

The linguistic representation consists of 36 syllable- and phrase-level features listed in Table 1. The syllable-level features referring to pitch accents and boundary tones are derived from the description provided by models performing automatic detection and classification of pitch accents and boundary tones. Information on the break index is derived partly from the annotation and partly from the information provided in the automatic classification of boundary tones. Other features can be easily derived from utterance's annotation. For each syllable the features of two previous and two following syllables are provided.

Phrase-level features are derived from the phrasing description provided by models trained to detect phrase boundaries and classify boundary types.

Table 1: *Syllable- (italics) and phrase-level features.*

Feature	value
stress	*stressed, unstressed*
accent	*unaccented,* *accented: LH*,L*H, HL*, H*L, LH*L*
boundary tone	*no boundary,* *boundary: 2,?, 2,., 5,?, 5,., 5,!*
break index	*0-4*
position	*distance to phrase start/end measured in the number of syll., stressed and accented syll.*
	distance to the start/end of the word (in syll.)
	distance to the next and previous pause in syllables and milliseconds
other	*length of the current word (in syll.)*
	syllable structure
	relative syllable/nucleus duration
phrase type	single (major int. phrase containing one minor phrase), initial, medial, final (see [18])
phrase length	expressed in the number of syllables
tune type	statement, exclamat., question, continuation

ANOVA results revealed that all the syllable- and phrase-level features have statistically significant effect ($p<0.01$) on pitch target scaling. The greatest effect was observed for features related to syllable position: distance to the start and end of the phrase measured in the number of pitch accented syllables and distance to the next accented syllable. It was also observed that all predictor features caused greater variation in the level of f0mid and f0end targets than in f0start target.

1.4.2. Duration estimation

Duration estimation was based on a set of 57 linguistic features. The choice of the features was aimed at recognizing, describing and assessing the dependencies of the segmental duration on those features and their combinations. It was intended to create a wide feature space and to test the dependencies for both large and smaller subsets of the feature set. The feature space included the most basic, segmental level features as well as features corresponding to higher levels of linguistic representation. For each speech sound a 7-element frame was used as the context information, i. e. the same properties were used as features for three preceding and for three following phones as well as for the phone in question.

The influence of the above features on duration was first tested for individual features by examining the mean duration distributions and ANOVA results which showed that the features connected with the position of the sound within higher level structures and the properties of the structures appeared to be statistically significant ($p<0.01$). The same was true for the word stress, and even more for the sentence stress (however since most often it falls on the penultimate syllable in Polish, both stress and accent usually coincide with the word-finality effect which might weaken the actual influence of the stress/accent factor). The effect of the word stress appeared to

be correlated with word length effect (the shorter the word the longer the component phones – this word length effect was stronger for the stressed phones, especially vowels).

Table 1: *Features for segmental duration prediction*

Feature	Feature description, values
Sound's Id	which particular Polish phone is the phone in question (39 values including palatalized [k g])
Sound's properties.	manner of articulation, place of articulation, presence of voice, type of sound (consonant or vowel); the same or different place of articulation of the phone in question and the phone in its direct left or right context
Properties of the preceding and of the following context.	ditto (for each of the phones within the 7-element frame)
	the information if the phone in question occurred in the neighborhood of an identical phone in the directly preceding or following context within the word in question
Position within the higher unit of speech organization structure	position of the sound in question relative to pause (3 categories)
	the distance of the syllable containing the phone to the left and right word boundary
	the position of the syllable within the foot (in the anacrusis, head or tail of the foot)
	the position of the foot within the intonation phrase (counted in feet)
	Position of the sound within syllable structure (onset, nucleus, coda)
Position relative to consonant clusters	Sound position within cluster, before or after cluster or with no cluster in the direct neighborhood
Speech unit length	Word length (in sounds, milisec.)
	Foot length (in sounds, in syll, milisec.)
	Syllable length (in sounds, milisec.)
	Phrase length (in sounds, in syll., milisec.)
Word stress	stressed, unstressed
Sentence stress.	accented, unaccented

3. F0 modeling

1.5. F0 target estimation

A neural network designer available in *Statistica 6.0* was used to create regression networks. MLP networks were selected for the preliminary training (using back propagation and/or conjugate gradient descend method) and assessment. All MLP networks had 3 layers and the number of units in the hidden layer varied between 1 and 20. Firstly, the networks were trained using the whole set of 36 predictor variables and then using an automatically selected feature subset. The databases containing neutral (N) and expressive (E) speech were divided into three subsets: training (N: 6011 syllables, E: 6892), selection (used among other things to halt or mitigate over-learning, N: 6019, E: 1430) and test (N: 5451, E: 1430).

In neutral speech, the best-performing MLP network (14 neurons in the hidden layer) relied on the whole set of the linguistic features (36). In expressive speech, the best results were achieved with an MLP network (10 hidden neurons) which used 15 input features selected automatically during the training.

Generally, regression results are better in neutral than in expressive speech. As regards correlation (r) between the original and estimated F0 targets it varies between **0.67** and **0.77** (test subset, overall(r)=**0.74**) in neutral speech and **0.63** and **0.7** in expressive speech (test subset, overall(r)=**0.68**). Low overall RMSE: N=11.33, E=0.72 (computed on the basis of z-scored normalized values) also indicates the accuracy of target estimation. The models designed on the basis of a Polish speech corpus and described in [6] have comparable performance in terms of correlation (between 0.68 and 0.75), but yield significantly higher RMSE (from 40 to 42Hz).

The results of sensitivity analyses carried out on the inputs to the regression networks revealed that **prosodic features** are the most important for an accurate F0 target level estimation. In the neutral speech data the key features were those related to pitch accents, boundary tones (e.g. the type of the pitch accent and/or boundary tone associated with the previous/ current/next syllable) and phrases (position in the utterance: initial/medial/final/single phrase; tune type). In the expressive speech data, 13 out of 15 input features were strictly prosodic (pitch accent or boundary tone type, strength of the previous/next prosodic break) or prosody-related (e.g. syllable position in the foot or phrase).

1.6. F0 contour generation

F0 contours for 98 utterances (32 neutral and 66 expressive) were generated using the procedure described in sec. 1.1.1 and subject to an objective and subjective evaluation.

1.6.1. Objective evaluation

RMSE and correlation coefficient (r) between the original and generated F0 contours were compared to assess objectively the similarity between the contours. It was observed that the proposed framework enables synthesis of F0 contours similar to the original contours irrespective of the speech style: if we average the results obtained in the expressive speech we get a correlation coefficient nearly the same as in the neutral speech: E=**0.73** vs. N=**0.72**, but higher average RMSE: E=**25.61** vs. N=**12.33**. These results compare favorably with [3] and [7], however in order to perform a more detailed comparison careful analysis of another factors influencing the results (such as the F0 variance) might appear necessary.

1.6.2. Perceptual evaluation

Two perception tests were carried out: similarity and quality tests. The results of the former showed no statistically significant differences with respect to speech style in the perceived similarity between stimuli with the original and synthesized intonation. In both neutral and expressive speech the stimuli were assessed as in between similar and a little different: 1.7<MOS<1.78 (MOS indicates the mean of all scores given by 10 subjects participating in the test, and 0 reflects no audible difference in intonation and 4 reflects no audible similarity). The quality of synthesized intonation was assessed on a 5-point scale (where -2 indicates *bad* and +2 - an *excellent* quality) and received scores between good (mean MOS(N)=0.17) and fair (mean MOS(E)=0.88). Almost 80% of the synthesized intonation contours from the neutral speech corpus and more than 50% contours from the expressive speech corpus assessed as perceptually different appeared to be functionally equivalent to the original intonation contours. Moreover, the overall quality of the synthesized intonation contours perceived as functionally different was rated as good

(mean MOS=0.3). The results of the perception tests confirm the results of objective evaluation i.e. similarity between the original and generated contours, and prove that the framework proposed in the current study can be used to generate high-quality, comprehensible and natural F0 contours.

4. Duration modeling

To obtain the duration model implemented later into BOSS, the CART implementation "wagon" was used for predicting the correlation between segmental duration and the features listed in Table 2 above. The correlation was calculated for the whole feature set (stepwise), and additionally, individual correlations for each particular feature were checked.

Detailed description of modeling Polish segmental duration with CART and the complete ranking lists of features were reported e.g. in [19]. When all 57 features were used for prediction, the overall correlation of the feature set was **0.8**, and the RMSE of **15.4**, which is relatively good as compared to other languages (e.g. [12, 20]), especially the error rate works out well in the comparison.

When comparing the stepwise correlation and the individual correlations of particular features it was found that two top-ranked features were the same in both sets. These were: the identity of the sound in question and the identity of the sound in the immediate following context. Another features were related to various properties of the sound's following context. The first higher-level features in both rankings were features corresponding to the position of the rhythmic foot containing the sound in question within the intonation phrase, and the distance of the foot from the end of the phrase. Another comparably significant features were word stress and sentence stress, and also the properties of the parent syllable.

The influence of expressiveness on segmental duration was not reflected in the present duration model, since it was not proved to be statistically significant in the present data [15]. However, some repeatable (and foreseeable) tendencies were observed for the text passage read by 40 speakers (segmental lengthening of the emphasized dialogue utterances or the negative correlation with the speech rate). The relationships observed for the regular speech material were generally confirmed to be true also for the 40-speaker speech sample, but, again, their statistical significance was not confirmed. The reason for the lack of the statistical significance of the features connected with speech expressiveness was most probably the humbleness of the speech data. In order to verify the dependencies in a more definite way it would be necessary to repeat the experiment using more appropriate speech signals.

5. Discussion and outlook

The most important quality of the proposed models is their applicability for speech technology. The linguistic representation is quite complex, since it is aimed at reflecting the diversity of connections between various levels of analysis as closely as possible. However, the resulting models are quite easy to implement, are characterized by low computational load and automated data-driven training without the need of extensive hand-tuning, and may be easily modified if needed. The results from the F0 model evaluation have shown that within the proposed framework high-quality, comprehensible and natural F0 patterns can be generated from a linguistically motivated representation in which prosodic features (e.g. lexical and sentence stress, the type of the pitch accent and/or boundary tone associated with the previous/current/next syllable) and positional features (e.g. syllable/foot position in the phrase, phrase position in the utterance) play a primary role. The duration prediction feature vector has been successfully implemented into the Polish speech synthesis in BOSS as one of the modules aiming at minimizing the system's cost function (unit-intrinsic costs). The proposed framework enables generating F0 and duration models, and the two types of models should complement each other, which is expected to be the initial step towards creating a common speech style independent prosody model for Polish. Obviously, in order to formulate more uniform and coherent framework, standardization of both models may be required.

6. References

[1] Pierrehumbert J.: The Phonology and Phonetics of English Intonation. PhD dissertation, MIT. [IULC edition, 1987]

[2] Taylor, P.: Analysis and synthesis of intonation using the tilt model. J. Acoust. Soc. Am 107(3):1697-1714 (2000)

[3] Möhler, G.: Improvements of the PaIntE model for F0 parameterization. Technical report. IMS, University of Stuttgart, September 2001

[4] Fujisaki, H. and Hirose, K.: Modeling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. Proc. 13th International Congress of Linguistics, pp.57-50, (1982)

[5] Black, A.W. and Hunt, A.J.: Generating F0 contours from ToBI labels using linear regression. Proc. of ICSLP 96', Philadelphia, USA, 1996, (3):1385-1388

[6] Oliver, D. and Clark, R.: Modeling pitch accent types for Polish speech synthesis. Proc. of INTERSPEECH 2005, Lisbon, Portugal, 2005, pp. 1965-1968.

[7] Dusterhoff, K., Black, A.W. and Taylor, P.: Using decision trees within the Tilt intonation model to predict f0 contours. Proc. of Eurospeech 99', Budapest, Hungary, 1999, pp. 1627–1630,

[8] Tesser, F., Cosi, P., Drioli, C. and Tisato, G.: Emotional Festival-Mbrola TTS synthesis. Proc. of INTERSPEECH 2005, Lisbon, Portugal, 2005, pp. 505-508

[9] Klatt, D. H.: Linguistic uses of segmental duration in English. Acoustic and perceptual evidence. JASA 59 (5), 1976, pp. 1208 – 1221

[10] Olaszy, G.: Predicting Hungarian Sound Durations for Continuous Speech. Acta Linguistica Hungarica, Budapest, vol. 49 (3-4):321-345 (2002)

[11] Riedi, M.P.: Controlling segmental duration in speech synthesis systems. PhD thesis, TIK-Schriftenreihe (26), ETH Zürich, 1998.

[12] Batusek, R. A.: Duration Model for Czech Text-to-Speech Synthesis. Proc. of Speech Prosody, Aix-en-Provence, 2002

[13] Campbell, N.: Multi-level timing in speech University of Sussex . PhD Thesis. (Exp. Psychol): Brighton, UK, 1992

[14] V. Santen, J.P.H.: Quantitative Modeling of Segmental Duration. Proc. of Human Language Technology Conference, Princeton, New Jersey, 1993, pp. 323-328

[15] Moebius, B. and van Santen, J.P.H.: Modeling segmental duration in German text-to-speech synthesis. Proc. of ICSLP, Philadelphia, PA, 1996, pp. 2395-2398.

[16] Richter, L.: The duration of Polish vowels. Speech Analysis and Synthesis, (3): 87-115., Warszawa, 1973

[17] Jassem, W., Krzyśko, M. and Stolarski, P.: Regression model of isochrony in speech signal, IPPT PAN, vol. 33/1981, Warszawa

[18] Clark, R.: Generating Synthetic Pitch Contours Using Prosodic Structure. PhD thesis, The University of Edinburgh, 2003.

[19] Szymański M. and Grocholewski S., Transcription-based automatic segmentation of speech. [in:] Proceedings of 2nd Language & Technology Conference, Poznań, 2005, pp. 11-15.

[20] Klessa, K., Szymański, M., Breuer, S.and Demenko, G.: Optimization of Polish Segmental Duration Prediction with CART., SSW6, Bonn, Germany 2007, pp.77-80

[21] Chung, H. and Huckvale, M.: Linguistic factors affecting timing in Korean with application to speech synthesis. Proc. of Eurospeech 2001, Aalborg, Denmark, 2001, pp. 815-818