A Comparison of Two Prosody Modelling Approaches for Sesotho and Serbian

Lehlohonolo Mohasi[1], Milan Sečujski[2], Robert Mak[2], Thomas Niesler[1]

[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
{lmohasi, trn}@sun.ac.za
[2]Faculty of Technical Sciences, University of Novi Sad, Serbia
secujski@uns.ac.rs

Abstract. Accurate prediction of prosodic features is one of the critical tasks within a text-to-speech system, especially for under-resourced languages with complex lexical prosody. For synthesized speech to have a natural-sounding intonational contour, an adequate prosodic model should be employed. This study compares the Fujisaki model and the HMM-based prosodic modeling in the context of text-to-speech synthesis, for two quite distant languages with rich prosodic systems: Sesotho, a tonal language from the Bantu family, and Serbian, a South-Slavic language with pitch accent. The results of our experiments suggest that, for both languages, the Fujisaki model outperforms the HMM-based model in the modelling of the intonation contours of utterances of human speech.

Keywords: prosody modelling, Fujisaki model, hidden Markov models, text-to-speech synthesis, Sesotho language, Serbian language.

1 Introduction

Accurate prosodic modelling is a crucial factor in order for text-to-speech (TTS) systems to produce intelligible and natural-sounding speech. Prosodic features include the fundamental frequency (F0) contour, duration, pause and amplitude. Tone, on the other hand, is a linguistic property marked by prosodic features such as F0 and intensity. Due to the absence of prosodic marking in the written format, automatic prosody generation for text-to-speech is a challenge for most languages, particularly those with more complex lexical prosody. This holds for both tonal languages such as Sesotho [1], as well as pitch-accent languages such as Serbian [2]. In this paper, we investigate and compare two tools which comprise prosody modelling and automatic prosody generation for text-to-speech systems, and calculate their efficiencies for the two languages mentioned above. In this research we focus on the modelling and prediction of F0, which is perceptually the most important element of sentence prosody.

The first method employs the Fujisaki model [3], which is reliant on the acoustics of the uttered speech. The Fujisaki model is a manageable and powerful model for prosody manipulation. It has shown a remarkable effectiveness in modelling the F0 contours and its validity has been tested for several languages, including tonal lan-

guages such as Mandarin [4] and Thai [5]. The second method, widely used within hidden Markov model based speech synthesis (HTS), employs a set of trained statistical models (context-dependent HMMs), which are used to predict prosodic parameters such as durations of particular phonetic segments and values of log F0. Both models rely on a previously recorded speech corpus, used to train the model, i.e. to set the values of its relevant parameters.

Section 2 gives a brief background of Sesotho and Serbian, followed by the description of the two models, the Fujisaki model (Section 3) and HTS (Section 4). Preparation of data for the experiments and the subsequent surface tone transcription are explained in Sections 5 and 6 respectively. Experimental results are illustrated in Section 7 and the conclusions drawn thereof are in Section 8.

2 Background on the Sesotho and Serbian Languages

Sesotho is a Southern Bantu tonal language spoken in Lesotho as a national language and in South Africa as one of the eleven official languages. It has two tone levels: high and low, of which the high tone is the active tone. Sesotho is classified as a grammatical tone language, which means that words may be pronounced with varying tonal patterns depending on their particular function in a sentence. In order to create certain grammatical constructs, tonal rules may modify the underlying tones of the word and thus lead to differing surface tones. The underlying tone, also known as the lexical tone, is the tonal pattern of the word in isolation. The surface tone, on the other hand, is a 'spoken' tone, i.e. the tone given to a word when spoken as part of a sentence. The surface tone can be derived from the underlying tone using tonal rules.

Serbian is the standardized variety of the Serbo-Croatian language, spoken in Serbia as the official language, and in some other countries of the Balkan peninsula as well. Serbo-Croatian is the only Slavic language which uses a pitch accent, assigning it at the level of the lexicon and using it to differentiate between word meanings or values of morphological categories. Traditional grammars define the pitch accent of Serbo-Croatian through four distinct accent types, which involve a rise or fall in pitch associated with either long or short vowels, and with optional post-accent lengths. However, more recent analyses ([6], [7]) have shown that these accent types can be interpreted as tonal sequences, i.e. reduced to sequences of high and low tones, without loss of representativeness, provided that phonemic length contrast is preserved. Thus, words can be thought of as strings of syllables following tonal patterns, and the surface tone of the utterance can be derived from its underlying tone using appropriate tonal rules.

3 The Fujisaki Model

The Fujisaki model, which is formulated in the log F0 domain, analyses the F0 contour of a natural utterance and decomposes it into a set of basic components which, together, lead to the F0 contour that closely resembles the original. The components are: a base frequency, a phrase component, which captures slower changes in the F0

contour as associated with intonation phrases, and a tone component that reflects faster changes in F0 associated with high tones. The tone commands of the Fujisaki analysis are an indicator of tones in the utterance.

The method was first proposed by Fujisaki and his co-workers in the 70s and 80s [8] as an analytical model which describes fundamental frequency variations in human speech. By design, it captures the essential mechanisms involved in speech production that are responsible for prosodic structure. A chief attraction of the Fujisaki model lies in its ability to offer a physiological interpretation that connects F0 movements with the dynamics of the larynx, a viewpoint not inherent in other currently-used intonation models which mainly aim to break down a given F0 contour into a sequence of 'shapes' [9]. The Fujisaki model has been integrated into a German TTS system and proved to produce high naturalness when compared with other approaches [10]. The inverse model, automated by Mixdorff [3], determines the Fujisaki parameters which best model the F0 contour. However, the representation of the F0 contour is not unique. In fact, the F0 contour can be approximated by the output of the model with arbitrary accuracy if an arbitrary number of commands is allowed [11]. Therefore, there is always a trade-off between minimizing the approximation error and obtaining a set of linguistically meaningful commands.

Mixdorff et al. [12] and Mohasi et al. [13] found that for Sesotho, the Fujisaki captures tone commands of positive amplitudes for the high tones. For other tonal languages that have been investigated using this technique, such as Mandarin [4], Thai [5], and Vietnamese [14], low tones are captured by tone commands of negative polarity. In contrast, low tones in Sesotho were found to be associated with the absence of tone commands. It should be noted that, unlike Sesotho, so far there has been no reported research into the modelling of intonation using the Fujisaki model for Serbian.

4 Hidden Markov Model-based Speech Synthesis (HTS)

HTS has been demonstrated to be very effective in synthesizing speech. The main advantage of this approach is its flexibility in changing speaker identities, emotions, and speaking styles. Statistical parametric speech synthesis using a hidden Markov model (HMM) as its generative model is typically called HMM-based speech synthesis. In this approach, HMMs represent not only the phoneme sequences but also various contexts of the linguistic specification. The models are trained on a speech corpus in order to be able to predict the values of prosodic and spectral parameters of speech for a given text. However, since it is impossible to prepare training data for all conceivable linguistic contexts, a number of tree-based clustering techniques have been proposed in order to allow HMMs to share model parameters among states in each cluster.

The synthesis part of the system converts a given text to be synthesized into a sequence of context-dependent labels. According to the label sequence, a sentence-level HMM is constructed by concatenating context-dependent HMMs. The duration of each state is determined to maximise its probability based on its state duration proba-

bility distribution. Then a sequence of speech parameters including spectral and excitation parameters is determined so as to maximise the HMM likelihood. Finally, a speech waveform is resynthesised directly from the generated spectral and excitation parameters by using a speech synthesis filter, for example, a mel-log spectral approximation filter for mel-cepstral coefficients or an all-pole filter for linear-prediction-based spectral parameter coefficients [15].

5 Data Preparation

The data used for the Sesotho corpus is based on a set of weather forecast bulletins obtained from the weather bureau in Lesotho, Lesotho Meteorological Services (LMS). The original data was compiled and broadcast for Lesotho TV. The original audio data was not of high quality, containing considerable background noise, as well as a large variability in speaking rate. The poor signal-to-noise ratio (SNR) in particular made this data unsuitable for eventual use in TTS development. For this reason, the sentences were re-recorded by the first author, who is a female native speaker of Sesotho. The recordings were performed in a quiet studio environment, at a sampling rate of 48 kHz. The corpus contains 40 minutes of speech and utterances are 12 seconds long on average.

The available Serbian speech corpus contains approximately 4 hours of speech, recorded in a sound-proof studio and sampled at 44 kHz. All sentences were uttered by a single female voice talent, a professional radio announcer using the ekavian standard pronunciation. General intonation in the database ranged from neutral to moderately expressive, and the effort was made to keep the speech rate approximately constant. However, in order to avoid a considerable difference in the experiment setup for Serbian and Sesotho, only a portion of the corpus corresponding in size to the entire Sesotho corpus was used for the principal experiment involving the comparison between the Fujisaki model and HMM prosody generation for both languages.

6 Surface Tone Transcription

Various studies show that the pronunciation of Sesotho lexical items can differ from one syntactic context to another. Words in isolation are pronounced differently from words in context in many languages, and therefore, there is a need to perform a surface tone transcription as part of the compilation of the corpus. The requirements for surface tone transcription are a pronunciation dictionary, a set of tonal rules, and a morphological analysis. The pronunciation dictionary for Sesotho is based on two tone-marked dictionaries, that by Du Plessis et al. [16] and Kriel et al. [17]. We used Sesotho tonal rules due to Khoali [18], whose work on Sesotho tone is the most recent. The sentences in the corpus were then annotated with underlying tones from the pronunciation dictionary, from which a surface tone transcription was deduced by means of a morphological analysis as well as tonal rules.

The sequence of tones for the Serbian corpus was determined based on the pitch accent assigned by the system for automatic POS tagging [19], with tagging errors

manually corrected. Appropriate tonal rules were used to convert the underlying accent to the surface accent (or, alternatively, to convert the underlying tone to surface tone).

Once the surface tone transcription was complete for both languages, the sentences were annotated at syllable levels using Praat TextGrid editor [20]. F0 values were extracted using Praat at a step of 10ms and inspected for errors. The F0 tracks were subsequently decomposed into their Fujisaki components applying an automatic method originally developed for German [21]. Initial experiments in [13] for Sesotho have shown that the low tones in the critical words of the minimal pairs could be modelled with sufficient accuracy without employing negative tone commands. Serbian, as illustrated in Figure 1, also shows positive tone commands only. As a consequence, only high tones were associated with tone commands. Adopting this rationale, automatically calculated parameters were viewed in the FujiParaEditor [22] and corrected when necessary.

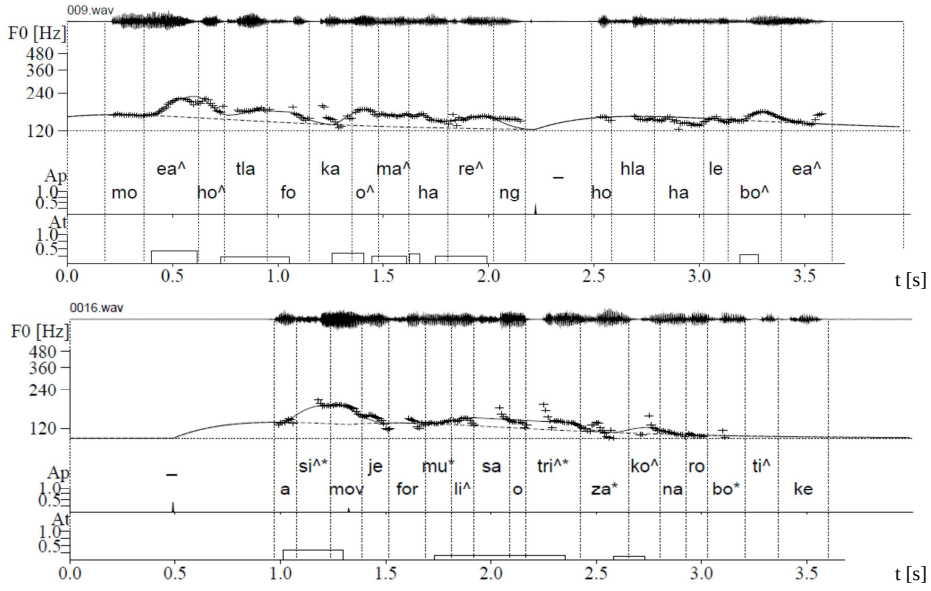

Fig. 1. Sesotho (009.wav) and Serbian (0016.wav) sentences illustrating the Fujisaki-modelled F0 contours (with crosses '+') and their surface tone labels. The high surface tones are indicated by the symbol '^', while stress is indicated by '*' for Serbian. Vertical dotted lines mark syllable boundaries. The Sesotho sentence reads "*Moea ho tla foka o mahareng, ho hlaha leboea.*" – "A moderate wind will blow from the north." and that for Serbian reads "*Asimov je formulisao tri zakona robotike.*" – "Asimov formulated the three laws of robotics."

7 Experiment Results

Ten utterances from each language were selected for testing by both models. The resulting synthesized data were then compared with original utterances, where root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (CC) were calculated for duration and pitch. Tables 1, 2, and 3 show the results obtained from these calculations.

Table 1. Comparison between original utterances and those resynthesized by a Fujisaki model.

	Duration (ms)		Pitch (Hz)	
	Sesotho	Serbian	Sesotho	Serbian
RMSE	0.00	0.00	9.18	14.41
MAE	0.00	0.00	6.28	8.38
CC	1.00	1.00	0.89	0.63

Table 2. Comparison between original utterances and those synthesized by a HTS system.

	Duration (ms)		Pitch (Hz)	
	Sesotho	Serbian	Sesotho	Serbian
RMSE	35.76	22.39	23.41	16.10
MAE	28.14	17.27	18.65	12.28
CC	0.27	0.74	0.03	0.59

Table 3. Comparison between utterances resynthesized by a Fujisaki model and those synthesized by a HTS system.

	Duration (ms)		Pitch (Hz)	
	Sesotho	Serbian	Sesotho	Serbian
RMSE	35.76	22.39	23.75	17.54
MAE	28.14	17.27	19.15	14.57
CC	0.27	0.74	-0.04	0.37

In Table 1, both languages show a positive high correlation for both duration and pitch, with Sesotho presenting a closer relationship than Serbian. Duration displays perfect correlation of the value of 1 for the two languages, which suggests that the duration of the utterances was not affected during resynthesis. This is also an indication that the F0 extracted by the Fujisaki model is closely similar to that of original utterances. The RMSE and MAE values do not show any variation for duration, and the variation in pitch is quite small.

In the comparison between the original utterances and those synthesized via a HTS system [23], given in Table 2, the values here demonstrate a substantial difference from those found in Table 1. Correlation for Serbian is significant for both duration and pitch, while that for Sesotho is quite low for both instances, with almost no correlation for pitch. The RMSE and MAE variations have increased for both languages, considerably so for duration. For pitch, the increase in variation is by a small margin for Serbian, while that for Sesotho is almost three times as much.

Table 3 compares the two tools and the scores attained are the same as in Table 2 for duration. For pitch, the increase in variation values is small in comparison to values obtained in Table 2. However, correlation has decreased for both languages, with Sesotho showing a negative correlation.

In general, the Fujisaki model has a more accurate F0 modelling capability than HTS. Although the Fujisaki model performed better for Sesotho, HTS showed a significantly higher performance for Serbian. This is due to the fact that Serbian had more data available for training in HTS, whereas data for Sesotho did not meet the minimum requirements.

8 Conclusion

In this paper we have explored and compared the F0 modelling capabilities of the Fujisaki model and HTS for two distant languages, Sesotho and Serbian. Accurate F0 (prosody) modelling is crucial for a natural-sounding text-to-speech system. The results obtained show the Fujisaki model to be more accurate than HTS. The prosodic-modelling accuracy of the HTS system can be improved by training more data, preferably more than 4 hours of speech. This is our next step, especially for the Sesotho language.

Acknowledgments: The presented study was sponsored in part by the National Research Foundation of the Republic of South Africa (grant UID 71926), by Telkom South Africa, and by the Ministry of Education and Science of the Republic of Serbia (grant TR32035).

References

1. Zerbian, S., Barnard, E.: Word-level prosody in Sotho-Tswana. In: Speech Prosody (2010).
2. Sečujski, M., Obradović, R., Pekar, D., Jovanov, Lj., Delić, V.: AlfaNum System for Speech Synthesis in Serbian Language. In: 5th Conf. Text, Speech and Dialogue, pp. 8-16 (2002)
3. Mixdorff, H.: A novel approach to the fully automatic extraction of Fujisaki model parameters. In: IEEE Int. Conference on Acoustics, Speech, and Signal Processing. Istanbul, Turkey, pp. 1281-1284 (2000)
4. Mixdorff, H., Fujisaki, H., Chen, G., Hu, Y.: Towards the automatic extraction of Fujisaki model parameters for Mandarin. In: Eurospeech/Interspeech 2003, pp. 873-876 (2003)
5. Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H.: Perception of tone and vowel quantity in Thai. In: ICSLP (2002)
6. Gođevac, S.: Transcribing Serbo-Croatian Intonation. In: S.-A. Jun (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 146-171, Oxford Linguistics, UK (2005)
7. Sečujski M., Jakovljević N., Pekar D.: Automatic Prosody Generation for Serbo-Croatian Speech Synthesis Based on Regression Trees. In: Interspeech 2011, pp. 3157-3160 (2011)

8. Fujisaki, H., Hirose, K.: Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Journal of the Acoustics Society of Japan (E) 5 (4), pp: 233-241 (1984)
9. Taylor, P.A.: The Rise/Fall/Connection model of intonation. In: Speech Communication Vol. 15, pp. 169-186 (1995)
10. Mixdorff, H., Mehnert, D.: Exploring the Naturalness of Several High-Quality Text-to-Speech Systems. In: Eurospeech 1999, pp. 1859-1862 (1999)
11. Aguero, P.D., Wimmer, K., Bonafonte, A.: Automatic analysis and synthesis of Fujisaki intonation model for TTS. In: Speech Prosody (2004)
12. Mixdorff, H., Mohasi, L., Machobane, M., Niesler, T.: A study on the perception of tone and intonation in Sesotho. In: Interspeech 2011, pp. 3181-3184 (2011)
13. Mohasi, L., Mixdorff, H., Niesler, T.: An acoustic analysis of tone in Sesotho. In: ICPhS XVII, pp. 17-21 (2011)
14. Dung, T.N., Luong, C.M., Vu, B.K., Mixdorff, H., Ngo, H.H.: Fujisaki model-based F0 contours in Vietnamese TTS. In: ICSLP (2004)
15. Masuko, T.: HMM-Based Speech Synthesis and Its Applications. Ph.D. thesis, Tokyo Institute of Technology, Japan (2002)
16. Du Plessis, J.A. et al.: Tweetalige Woordeboek Afrikaans-Suid-Sotho. Via Afrika Bpk, Kaapstad, SA (1974)
17. Kriel, T.J., van Wyk, E.B.; Pukuntsu Woordeboek Noord Sotho-Afrikaans, Van Schaik, Pretoria, SA (1989)
18. Khoali, B.T.: A Sesotho Tonal Grammar. PhD Thesis. University of Illinois, Urbana-Champaign, USA (1991)
19. Sečujski, M., Delić, V. A Software Tool for Semi-Automatic Part-of-Speech Tagging and Sentence Accentuation in Serbian Language. In IS-LTC (2006)
20. Boersma, P.: Praat - A system for doing phonetics by computer. In: Glot International, Vol. 5, No. 9/10, pp. 341-345 (2001)
21. Mixdorff, H.: Intonation Patterns of German - Model-based Quantitative Analysis and Synthesis of F0 Contours. PhD Thesis. TU Dresden, Germany (1998)
22. Mixdorff, H.: FujiParaEditor. http://public.beuth-hochschule.de/~mixdorff/thesis/ fujisaki.html (2012)
23. Pakoci, E., Mak, R.: HMM-based Speech Synthesis for the Serbian Language. In: ETRAN, Zlatibor, Serbia (2012)