Whispered Speech Prosody Modeling for TTS Synthesis

Valery A. Petrushin [1]*, Liliya I. Tsirulnik* [2] *and Veronika Makarova*[3]

[1] The Nielsen Company, Schaumburg, IL, USA
[2] Speech Recognition and Synthesis Laboratory,
United Institute of Informatics Problems of NAS of Belarus, Minsk, Belarus
[3] Department of Languages and Linguistics, University of Saskatchewan, Saskatoon, Canada
vapetr3@hotmail.com, liliya.tsirulnik@gmail.com, v.makarova@usask.ca

Abstract

This paper is devoted to modeling prosody of whispered Russian speech. The practical purpose of this research is to extend voice cloning techniques to whispered speech modality. The authors present their analysis of prosodic features that contribute to the expression of sentence type intonation in whispered speech. The current investigation includes intonation contours in complete and incomplete declaratives, as well as in interrogatives and exclamations. Since the fundamental frequency is absent in whisper, the major role in conveying sentence type intonation is taken over by formant values. For modeling prosody of whispered speech, an extension of the Accent Unit Portrait Model is proposed. The paper demonstrates how melodic, rhythmic and dynamic (energy) portraits of accent units can be built and employed for whispered speech modifications by a concatenative text-to-speech synthesizer.

Index Terms: whispered speech, prosody modeling, speech synthesis, accent unit portrait model, formant modification.

1. Introduction

Whisper is defined as an unvoiced mode of phonation in which the vocal cords do not vibrate, but are adducted sufficiently to create audible turbulence as the speaker exhales during speech [1]. Whispered speech production is found in most world languages and is considered to be one of linguistic universals [2, 3]. The social role of whispering is to communicate information to neaby listeners without being overheard by other people. Whisper is also used for communication by aphonic individuals who may be unable to produce vocal cord vibrations [4].

Over the last century, this speech modality has been examined within frameworks provided by a range of disciplines, such as speech science, phonetics, acoustics, engineering, medicine and health science. Studies of whispered speech find practical applications in evaluation of voice and hearing disorders, speaker and speech recognition for forensic, security, military and other purposes. A recent wave of interest in whispered speech research is explained by a wide spread of speech enabled devices such as mobile phones, smart phones and PDAs.

Below we shall use the term “regular” phonation (speech) to differentiate between whisper and the usual every day speech based on alternating voiced and voiceless segments.

From the speech production viewpoint, the major difference between regular and whispered speech is the state of the glottis. In regular speech, the vocal folds of the larynx vibrate and excite the resonances of the vocal tract. In whisper, the glottis is open, and the sound source is created by airflow passing though the glottis and reflected by the lower part of the vocal tract, which is about 2 cm in length and includes ventricular folds and epiglottis. As the result, whispered speech is noise excited, has power about 20 dB lower as compared to regular speech, and has no fundamental frequency component [4, 5, 6].

Some differences in the segmental quality of whispered vs. regular speech have been identified. The spectra in whispered speech have been reported to roll off under 500 Hz and be flatter than regular spectra between 500 and 2000 Hz [4, 5, 6]. Formant values (particularly F1 and F2) are known to shift in whisper [7, 8, 9]. While some studies show a general trend towards upward move of F1 and F2 in whisper, some other studies indicate that the exact direction and magnitude of the formant shift may be language or vowel specific. For example, in Serbian whispered vowels, F1 and F2 values were rising for /i ,e, a, o/ and dropping for /u/ [5].

One of the most intriguing aspects observed in speech studies is the ability of language speakers to trace pitch in whisper despite the absence of F0. The first explanation has been suggested by H. von Helmholtz in 1950s. Using tuning forks and whistles, he determined that in whisper, the perceived pitch of back vowels /u, o, a/ corresponds to the values of first formant (F1) 175, 466, 932 Hz, while the perceived pitch of front vowels /y, e, i/ corresponds to the values of second formant (F2) 1568, 1876, 2349 Hz [12]. Recently a number of more detailed research studies related to whispered speech perception and prosody has been conducted [9, 13, 14, 15]. Some other studies addressed the prosody of whisper in tone languages [16, 17].

One of the first studies devoted to signal processing of whispered speech including whisper-to-speech conversion, and speech and speaker recognition is [8]. The author has applied jump Markov linear systems to determine interframe relationships in the mixed excitation linear prediction (MELP) algorithm that was used for modeling and coding speech. A whisper-to-speech system and a MELP-based synthesizer have been created based on the above models [3, 4]. New methods have been subsequently developed and applied for modifying linear prediction spectra, shifting formants and synthesizing the excitation signal in the real-time reconstruction of regular speech from whisper [3].

2. Modeling prosody of Russian whispered speech

The objective of voice cloning technology is to model voices of individual speakers including all their acoustic, prosodic and lexical idiosyncrasies [16]. This project aims at improving voice cloning technology and extending it to whispered speech. The paper presents some preliminary results of intonation analysis in whispered speech and proposes a model for capturing speaker-dependent implementations of different types of sentence intonation. The proposed model is used for prosodic modification of speech units of the whispered speech synthesis system [17].

2.1. Speech data

For our analysis of sentence intonation in whispered speech, we recorded two sets of utterances pronounced by a male and a female speaker. The first set consists of meaningless 3-syllable one-word utterances with the CVCVCV structure, where C is the stop consonant /b/, and V is one of the five Russian vowel phonemes {/a/, /e/, /i/, /o/, /u/} used repeatedly in all the three syllables. Each utterance was recorded in regular and whispered modes with the lexical stress on the first, second and third syllable. All these accentual variants were produced with two different intonation contours: as a complete declarative sentence and as a Yes/No interrogative. Each of the resulting stimuli was recorded five times. Overall we obtained 300 utterances (2 speech modes x 2 intonation types x 5 vowels x 3 stress assignments x 5 repetitions=300) per speaker.

The second set of experimental stimuli consists of the following five meaningful Russian sentences:

1. “Брат пил чай [, кофе и по субботам коньяк.]” (/b/r/a/t/_/p’/i/l/_/ch’/a/j’[/_/k/o/f’/e/_/i/p/a/s/u/b/o/t/a/m/_/k/a/n’/j’/a/k], “The brother drank tea, [coffee and cognac on Saturdays]”.

2. “Дети строили башню, [замок и ров] (“D'/e/t’/i/_/s/t/r/o/i/l’/i/_/b/a/sh/n’/u [/_/z/a/m/o/k/_/i/r/o/f], “Kids built a tower, [a castle and a moat]”.

3. “В лесу водятся змеи, [волки и медведи]” (/v/l'/e/s/u/_/v/o/d’/a/t/s/a/_/z/m’/e/i[/_/v/o/l/k’/i/_/i/m’/e/d/v’/e/d’/i], “In the forest, there are snakes, [wolfs and bears]”.

4. “Машина работала хорошо [надёжно и бесшумно] (/m/a/sh/i/n/a/_/r/a/b/o/t/a/l/a/_/h/a/r/a/sh/o[/_/n/a/d’/o/zh/n/a/_/i/b’/e/s/sh/u/m/n/a], “The machine worked well, [reliably and noiselessly]”.

5. “Работа завершена вовремя, [быстро и качественно]” (/r/a/b/o/t/a/_/z/a/v’/e/r/sh/e/n/a/_/v/o/v/r’/e/m’/a[/_/b/y/s/t/r/a/_/i/k/a/ch’/e/s/t/v’/e/n/n/a], “The work is completed on time, [quickly and with quality]”.

Each utterance was read in whisper by a male speaker of Russian as a declarative, exclamation and an interrogative. To compare the intonation in complete (final) and incomplete (non-final) utterances, two versions of declarative utterances were recorded: including and excluding the text in brackets above. Only the intonation of the first phrase in each utterance (before the brackets in the text above) was analyzed. Altogether, the data set consists of 100 utterances (5 sentences x 4 intonation types x 5 repetitions)

Both speech data sets were manually segmented and labeled on segmental and prosodic (accent unit) tiers. The following features were extracted for acoustic analysis: formant values F1 and F2 for all vowels, energy and duration of each accent unit component.

2.2. The prosodic model

For modeling prosody of whispered speech, we propose to use the extension of the Lobanov-Karnevskaya prosodic model, which is based on British Tone Unit approach [18], and known as the Accent Unit Portrait (AUP) Model [19]. According to this model, each prosodic phrase consists of one or more accent units (AU). An AU has only one accented syllable, but may include one or more words. The AU is divided into three parts: *a nucleus (*the accented syllable), a *pre-nucleus (*unstressed syllables preceding the nucleus), and a *post-nucleus* (unstressed syllables following the nucleus). The major assumption of the AUP model is that the topological prosodic properties of a certain intonation pattern (tone) do not change significantly when the phonetic content and the number of syllables changes in the pre- and post-nuclei. Figure 1 shows F0 profiles for a three word question that is pronounced in regular speech three times with accent on the first, second and third word. The normalized profiles (portraits) for all three utterances look very similar. Creating portraits for different types of intonation with different number of AUs allows to capture both typological and individual realizations of intonation contours. The obtained portraits are subsequently used in a TTS synthesis system for simulating prosodic characteristics of a speaker. Usually the number of portraits ranges from 100 to 200. Each portrait includes profiles for pitch, energy and rhythm (duration).

2.3. Prosodic characteristics of regular and whispered speech

Table 1 and Figure 2 show the average values and standard deviations of F1 and F2 formants in the five regular and whispered Russian vowels extracted from the first data set for the male speaker. The formants of all whispered vowels are shifted to higher values. These results confirm the findings of previous research [5, 8].

Table 1. *F1 & F2 average values for normal and whispered speech.*

Vowel	F1 mean	F1 std	F2 mean	F2 std
/a/ voiced	682	19	1205	52
/a/ whispered	918	48	1542	90
/e/ voiced	546	16	1750	24
/e/ whispered	726	51	1850	67
/o/ voiced	456	38	748	40
/o/ whispered	632	73	902	53
/u/ voiced	325	25	667	46
/u/ whispered	523	46	723	39
/i/ voiced	351	41	2170	61
/i/ whispered	426	37	2275	38

Energy and duration values extracted from the utterances in both speech data sets show that energy of whispered speech is about 20-30 dB lower, but duration of whispered utterances is about 20% longer. Similar results were obtained in our earlier studies that showed decreased speech rate and increased vowel and consonant duration in whisper as compared to regular Russian speech [17].

2.4. Analysis of whispered speech prosody

Analysis of the first speech data set allows us to compare prosodic features of one-word (one accent unit) utterances across two sentence intonation types: declarative final and interrogative. The results indicate that F2 plays major role in expressing intonation in whisper. The values of F1 also change in whisper, but less significantly. Figure 3 shows F1 and F2 averages (normalized by duration and frequency profiles) across both intonation types. F2 values are significantly higher in interrogative utterances than in declaratives, whereby the difference is 20% in the accented syllable and 10% in the post-accented syllable. F1 values in accented syllables of the declarative utterances are insignificantly higher than in interrogatives, whereas F1 in the post-accented syllables is about 6% lower in declaratives than in interrogatives. Accented syllables are longer in interrogatives than in declaratives.

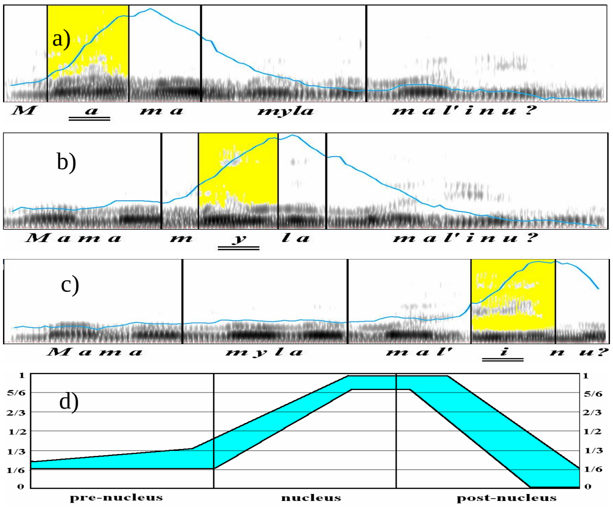

Figure 1. *Pitch contours for an interrogative phrase with the sentence accent on the first (a), second (b) and third (c) word and (d) its generalized melodic portrait.*

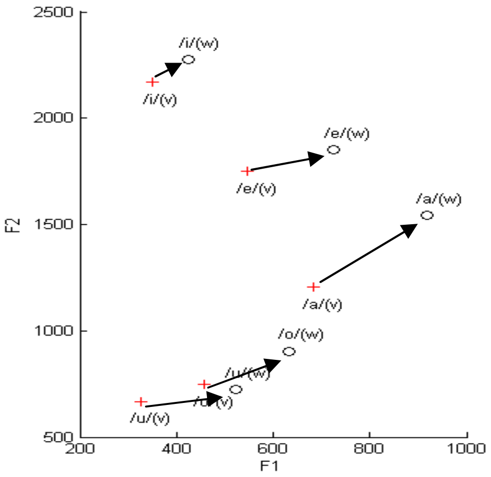

Figure 2. *Average values of F1 and F2 formants in regular and whispered speech.*

In interrogatives, the duration of the accented syllable increases with the shift of the lexical stress (and respectively accent) from the first to the last syllable in the word. The increase in the accented syllable duration in interrogatives as compared to declaratives is about 15% when the first syllable is stressed, 25% when the second syllable is stressed and about 40% when the third syllable is stressed.

Analysis of the second data set (containing three accent units) demonstrated some prosodic differences across the four intonation types: declarative final, declarative non-final, exclamatory and interrogative.

Unfortunately, the size of the data set was not large enough to make statistically significant conclusions. The following finding should be considered preliminary.

In most cases, F2 values are the highest in interrogatives, followed by exclamations and by non-final utterances, and are the lowest in final declaratives. The values of F1 are the highest in exclamation, followed by interrogatives, non-final and final declaratives. Duration values are approximately equal for final and non-final declaratives, while exclamatory phrases are about 10%-15% shorter and interrogative phrases are about 15%-20% longer than declaratives.

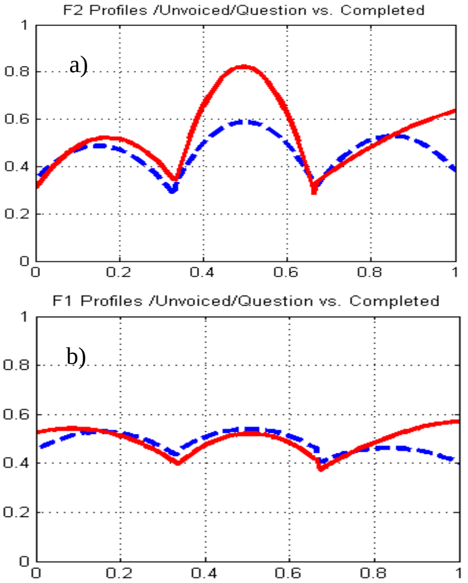

Figure 3. *Formant F2 (a) and F1 (b) profiles for declarative (dashed line) and interrogative (solid line) intonations.*

2.5. Modeling prosody of whispered speech

In regular speech, F0 functions as an acoustic correlate of pitch. In whisper, information about pitch is carried by the formants. To create a pitch profile for a whispered utterance, we have to create profiles for each formant. We chose to use the formants F1 and F2 in our model. Since each vowel has specific formant ranges, we have to model formant ranges for each vowel individually. Let us assume that $[FL_i^v, FU_i^v]$ is the range of *i*-th formant values for vowel *v*, where FL_i^v is the lower bound and FU_i^v is the upper bound for the vowel for a particular speaker. To create a pitch profile for *i*-th formant for a phrase of a specified prosodic type we use the following procedure.

1. Estimates lower and upper formant bounds for each vowel using all recordings of the speaker.
2. Collect all prosodic phrases of the specified type.
3. For each phrase do the following.
    a. Estimate the *i*-th formant values for each vowel in each phrase and normalize them

using the formant ranges:

$$F_{i\,norm}(t) = (F_i(t) - FL_i^v)/(FU_i^v - FL_i^v)$$

   b. Interpolate formant values in between the vowels using linear interpolation.
   c. Mark the boundaries of each AU splitting it into pre-nucleus, nucleus and post-nucleus.
   d. Normalize boundaries by duration by making all nuclei and pre-/post-nuclei of equal size.
   e. Estimate average formant bandwidths and keep their values with each part of AU.

4. Merge all profiles and estimate the average profile for *i*-th formant.

The whole portrait of the prosodic phrase of specified type consists of pitch, energy and rhythmic profiles (Figure 4). The details of estimating the energy and rhythmic profiles are omitted in this paper because we use the same techniques as the original AUP model [16].

The advantage of the AUP model is that it creates normalized templates for prosodic phrases which can be applied to speech units of the speech synthesis system for simulating phrase prosody. This process consists of several steps: first, the formant values of the speech unit are normalized and the differences between them and the portrait values are estimated, then the differences are converted into absolute values (shifts) and the formant modification algorithm is applied to the formant shifts and bandwidths using pole modification approach [20].

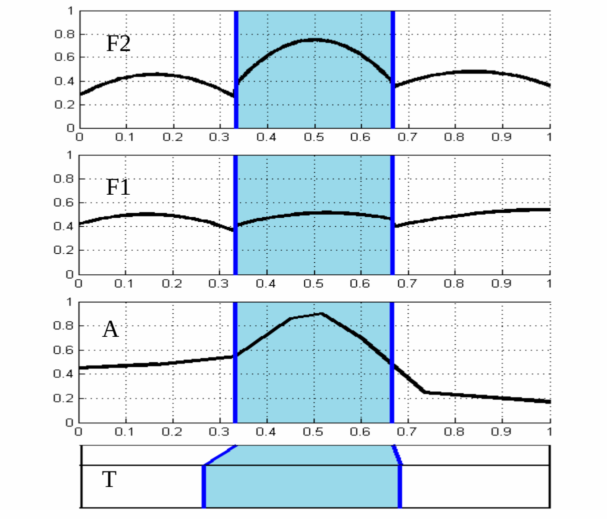

Figure 4. *The portrait of interrogative phrase with one accented unit. It consists of average profiles for F2 and F1 formants, energy (A), and rhythm (T).*

3. Conclusion and Future Work

In this paper, we made the first attempt to extent the AUP prosodic model to whispered speech. The complexity of the extended model is essentially higher than the complexity of the original one – it requires accounting for formants and their bandwidths, which are vowel specific. We realize that we made just one step toward developing whispered speech synthesis technology. In future, we plan to collect more data, build portraits for different intonation types and phrases of different accentual and segmental structure, improve our formant extraction and formant modification algorithms, and build a tool that helps creating prosodic portraits.

4. References

[1] Laver, J., “Principles of Phonetics”, Cambridge University Press, 1994.

[2] Rogers, H. “The Sounds of Language: An Introduction to Phonetics”, Harlow: Longman, 2000.

[3] Cirillo, J & Todt, D. “Perception and judgment of whispered vocalizations”, Behaviour, Vol 142, 2005, pp. 113-128

[4] Morris, R. W., Clements, M. A. “Reconstruction of Speech from Whispers”, Medical Engineering & Physics, Vol 24, 2002, pp 515-520.

[5] Jovicic, S.T., “Formant feature differences between whispered and voiced sustained vowels”, Acta Acoustica, vol. 84, no. 4, 1998, pp. 739-743.

[6] Schwartz, M. F. “Power spectral density measurements of oral and whispered speech”, J. Speech Hearing Res. Vol 13, 1970, pp 445-456.

[7] Tartter, V. C. “What’s in a whisper”, J Acoust. Soc. Am., Vol 86, 1989, pp 1678-1683.

[8] Morris, R.W.,” Enhancement and Recognition of Whispered Speech”, Ph.D. Dissertation, 2003, 159 p., accessed on 19 Nov 2009 , http://etd.gatech.edu/theses/available/etd-04082004-180338/unrestricted/morris_robert_w_200312_phd.pdf

[9] Higashikawa, M, and Minifie, F. D., “Acoustical-Perceptual Correlates of “Whisper Pitch” in Synthetically Generated Vowels”, J. Speech, Language, and Hearing Research, Vol. 42, 1999, pp. 583-591.

[10] von Helmholtz, H. L. F., “On the sensations of tone”, New York: Dover Publications, 1954.

[11] van Rossum, M., A., de Krom, G., Nooteboom, S., G., Quene, H. ““Pitch” Accent in Alaryngeal Speech”, J. Speech, Language, and Hearing Research, Vol. 45, 2002, pp. 1106-1118.

[12] van Rossum, M., A. “Prosody in Alaryngeal Speech”, PhD Dissertation, Utrecht: LOT, The Netherlands, 164 p.

[13] Kong, Y-Y., Zeng, F-G., “Temporal and spectral cues in Mandarin tone recognition”, J. Acoust. Soc. Amer., Vol. 120, 2006, pp. 2830- 2840.

[14] Chang, Ch., Yao, Y., “Tone production in whispered Mandarin”, UC Berkeley Phonology Lab Annual Report, 2007, pp. 326-329.

[15] Lin Wei, Yang Lili, Xu Boling. „A new frequency scale of Chinese whispered speech in the application of speaker identification”, Progress in Natural Science. Vol 16 (10), pp. 1072-1078.

[16] Lobanov, B., Tsirulnik L. “Phonetic-Acoustical Problems of Personal Voice Cloning by TTS”, Proc. SPECOM’2004, pp 17-21.

[17] Tsirulnik, L. I., Petrushin, V. A., Makarova, V. “Analysis and TTS-synthesis of Russian Whispered Speech”, Proc. SPECOM 2009, pp 180-185.

[18] Cruttenden, A. “Intonation”, Cambridge University Press. 1986.

[19] Lobanov, B. “Phonemophon Text-to-Speech System”, Proc. 11 Congress of Phonetic Sciences (ICPhS’87), Tallin, 1987, pp.61-64.

[20] Mizuno, H., Abe, M. “A formant frequency modification algorithm dealing with the pole interaction”, Electronics and Communication in Japan, Part 3, vol. 79, 1996, pp. 46-54.