From text to formants - indirect model for trajectory prediction based on a multi-speaker parallel speech database

Kálmán Abari[1], *Tamás Gábor Csapó*[2] , *Bálint Pál Tóth*[2], *Gábor Olaszy*[2]

[1] University of Debrecen, Hungary

[2] Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, Hungary

abari.kalman@arts.unideb.hu,{csapot,toth.b,olaszy}@tmit.bme.hu

Abstract

An indirect model is presented, capable of estimating formant trajectories from text only (Text-to-Formants, TTF). The result is a phonetically correct formant trajectory flow of any virtual speech signal, i.e. one that has never been uttered. The focus is on the pattern forms inside the given sound, taking into account the sound environment (up to quinphone), and not on individual formant value measurements. The model is based on a multi-speaker parallel speech database with precise manual corrections and a HMM-based formant trajectory predictor. The validation of the TTF model shows that formant trajectories can be predicted with good accuracy from text. The model indirectly gives information about a theoretically possible articulation flow of the sentence. Thus it gives a general 'formantprint' of the language.

Index Terms: formant trajectory prediction, HMM, reference formant database, sentence pattern, multi-speaker, parallel.

10.21437/Interspeech.2015-216

1. Introduction

An indirect Text-to-Formant (TTF) model for generation of formant patterns for the Hungarian language is introduced. The goal of this research is to demonstrate that formant values can be calculated not only from speech but also from a generative statistical model without making any measurements but using only the text. The input of the model is text, the output is the characteristic formant pattern flow of the sentence for F1 and F2. The model consists of two main parts: 1) precisely prepared multi-speaker parallel speech database with manually corrected sound boundaries and formant values (Formant Database, FDB); 2) HMM-based formant trajectory predictor from text. This is a novel indirect way of formant prediction from pure textual input. As the model gives indirect information about the articulation flow, effects of coarticulation can also be studied. The model may represent a language footprint ('formantprint') by connecting the textual content with associated formant trajectories. There are only a small number of public databases concerning the formant data of a given speech corpus. One of them contains the first three vocal tract resonances of 538 English sentences of TIMIT database, and is publicly available [1]. An Arabic formant database was used by Jemaa et al. [2] for the evaluation of a new automatic formant tracking algorithm based on Fourier ridges detection. A public online data inventory of formant maps has been published for Hungarian, which is based on isolated words of a male and a female speaker [3,4]. Work on a large-scale Hungarian formant database began recently [5], and the final version of it (FDB) is used in the current research. There are already many formant tracking algorithms available, working with speech signals. In our study, Snack [6,7], and Praat [8] were also used. The accuracy of formant trackers is generally speaker dependent [1]. An example of Praat formant tracking is shown in Fig. 1 for different speakers. F1, F2 and F3 values are denoted by blue, purple and green circles respectively. The output of the tracker is spoiled by mixed up values in all cases. Praat performed well for speaker (b) where only 2 mismeasured points occurred in F3. The trajectories of the three other examples (a, c, d) contain severe inaccuracies. This was the reason why the authors decided to correct the formant data in FDB manually.

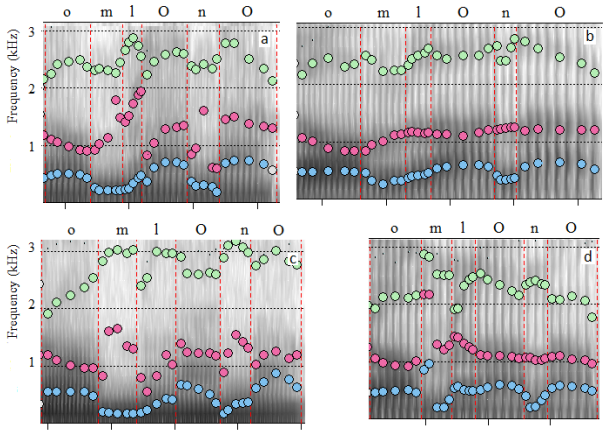

Figure 1: *Automatic formant tracking of Praat for the same sound group in the same sentence of 4 different speakers.*

In this paper similar principles are applied as in the case of HMM-based speech synthesis with the difference that the parameters to be modeled are formant trajectories, and we do not generate speech. State durations are also modeled by HMMs. This is different from an earlier work where formants were used as an intermediate representation for HMM-based speech synthesis [9].

2. Methods

2.1. Multi-speaker parallel speech database

For the development we used a multi-speaker parallel speech database [10], which was recorded at 44.1 kHz 16 bit in a professional studio. Parallel means that ten speakers (5 male and 5 female) read the same sentence corpus, i.e. 1,900

phonetically balanced sentences [11]. Altogether the speech corpus contained 10×1,900 sentences. Each sentence was automatically labelled and segmented. In the next step sound boundaries were visually controlled and manually corrected. SAMPA sound symbols were applied for the phonetic representation in the corpus. The formant estimation was done in two steps. First, automatic formant tracking was done by Praat, and in the second step the final formant values were manually adjusted by using a dedicated GUI tool [12]. Five points (10, 25, 50, 75 and 90%) inside every sound were determined to characterize the inherent formant pattern of the sound. Altogether 3×5=15 formant values represented each sound. The results are stored in the formant database (FDB). Two groups of speech sounds were defined concerning the formant measurements, i.e. those sounds that have characteristic formant structure (vowels and the consonants v, j, l, m, n and J; altogether 475,400 sounds in FDB) and those that do not (altogether 300,140). In the first group the formant values were defined from the speech signal (altogether 7,125,000 points); in the latter group only virtual data were defined by a linear interpolation between the adjacent measured data (altogether 4,502,100 points). The 5 point representation of formant patterns results in good enough formant movement description, clear and unified data structure and good opportunity for visual control, i.e. the manual corrections can be done clearly by moving the actual point vertically with the cursor (see section 2.4). Altogether 31.6% of the measured points of FDB had to be corrected manually. The higher the formant, the more corrections had to be made. In Figure 2, a sample sentence extract is shown from FDB with the corrected formant data (coloured circles) and with the automatically defined virtual data (white circles).

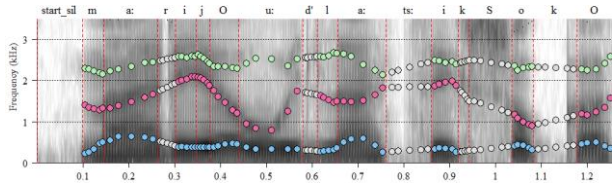

Figure 2: *A sentence extract from FDB with the 5 points/sound formant data*

The formant values are in tab delimited text files, where every sound of the sentence is represented by 5 rows according to the 5 points inside the sound.

Table 1 shows the set-up of the text file. Column *file* shows the identifier of the sentence and the speaker. *Num* identifies the position of the vowel within the speech sample: which segmented element of the line it is in. *Label* is the symbol of the vowel in SAMPA notation. *Time* is a point on the time axis (measured in seconds), which marks the place of the measurement. *Pos* is an identifier which differentiates points of measurement within a vowel (for example, 10 means the 10% measurement point). The last three columns list F1-F3 formant frequency values in Hz. This txt material represents the FDB (approx. 220MB).

Table 1. *A few lines from the formant database (FDB).*

file	label	num	time	pos	F1	F2	F3
fo1	o:	3	0.3893	10	401	1607	2246
fo1	o:	3	0.4096	25	440	1266	2267
fo1	o:	3	0.4435	50	456	1025	2445
fo1	o:	3	0.4774	75	457	1046	2602
fo1	o:	3	0.4978	90	282	1081	2425

2.2. A GUI tool for development and verification

We developed an application with graphical user interface for visualizing and correcting errors. The program called PROFEF is the improved version of the web based Interactive Formant Editor [13]. PROFEF supports both manual and automatic methods for error corrections. It has indexing features for accessing sentences that are the smallest elements on the screen. By indexing one can select and exclude sentences for analysis. The main window of PROFEF displays the spectrogram on which five points represent formant frequencies for F1, F2 and F3. These points are movable vertically to set the correct formant frequencies (Figure 3). The smallest step is 15.3 Hz/pixel. All changes are logged and can be undone.

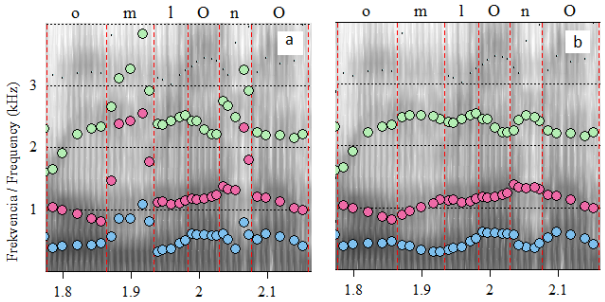

Figure 3: *Formants in a sentence extract from FDB before (a) and after manual correction (b). In m and n consonants 25 points were manually corrected*

2.3. HMM training

The FDB was divided into 2 parts: training database as corpus for training (90% of FDB) and verification database (VDB) as corpus for verification (10% of FDB). The training of the HMMs was done with the HTS toolkit [14] (version 2.2). F1 and F2 from FDB were trained because the goal was to build speaker independent (average) models and F3 is known to be strongly speaker dependent [15]. To be compatible with the HTS toolkit, the F1 and F2 formant trajectories of FDB were linearly interpolated to 5 ms intervals. These values are modelled with MSD-HMMs because they do not have values in silences. Logarithmic values are used as they were found to be more suitable in training experiments. The first and second derivatives of the parameters are also stored in the parameter files and are used in the training phase. Decision tree-based context clustering is used with context dependent labelling applied in the Hungarian version of HTS [16]. For the thorough evaluation of the general model we trained TTF models with various numbers of speakers as training data: models of single male speakers (altogether 5 models according to the 5 speakers, denoted by 1sp.m), single female speakers (5 models, denoted by 1sp.f, 5 male speakers (1 averaged model, denoted by 5sp.m) and 5 female speakers (1 averaged model, denoted by 5sp.f).

2.4. Concept of formant trajectory verification

To compare the formant movements produced by the TTF model the formant patterns of the sentences in the verification database were used, namely these were not used by the training procedures of the HMM models. Thus they are unknown for the TTF model. The process of comparison is as follows.

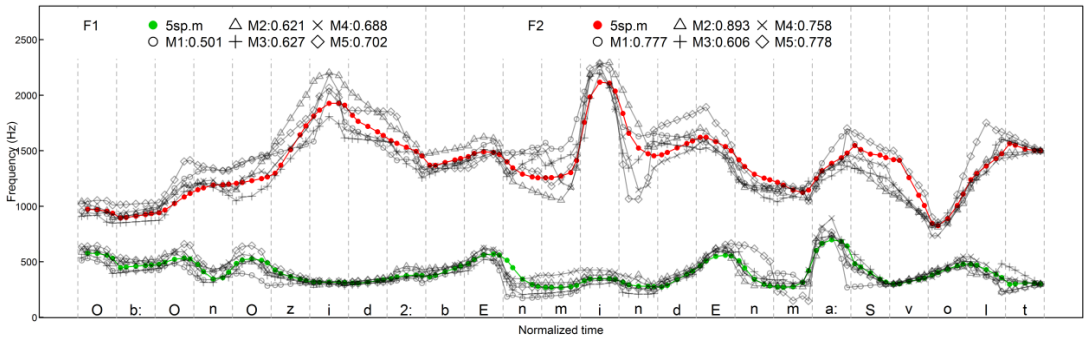

Figure 4: *TMR values for the vowels (in the legend) and the F1 and F2 formant contours for a sentence pronounced by the 5 male speakers (M1-M5) in comparison with the results of the TTF 5sp.m model. The F1 and F2 shapes of the sentence may be considered very similar to that of the individual speakers.*

The TTF model predicts the formant patterns from the text of the sentences of the verification database. These predicted data will be compared with the manually corrected formant patterns of the natural sentences of the 10 speakers of the verification database sentence by sentence. The procedure for comparison is called trajectory matching; it is partly self proposed. A new degree of the similarity between the predicted and the natural formant patterns is expressed by the Trajectory Matching Rate (TMR). It is a value between –1 and +1. The more similar the predicted formant pattern of TTF to that of the natural sentence in the verification database, the closer its TMR value is to +1. It is important that TMR does not show the equality degree of the formant values in Hz, only the similarity of patterns. TMR calculation is applied separately for F1 and F2. An important feature of this calculation is that it always concerns the same sentences. The predicted formant movements characterize the given sentence. The actual F1 and F2 curves represent the sentence. In general we refer to them as sentence patterns. Every sentence produced by the TTF model has its characteristic pattern for F1 and F2 independently of the speaker. An example is shown in Figure 4 where the predicted F1 and F2 sentence patterns of 5sp.m (green and red) are compared with the same natural data (5 speakers) in the same sentence. It can be seen that the general shape both for the F1 and F2 patterns is the same for every speaker; only individual pronunciation differences can be seen.

If all TMR values are averaged for all the sentences of the verification database, we get the general result showing how good the model is. This result expresses in general how similar the given predicted formant pattern of the TTF model is against the same natural sentences of the verification database.

The first step of TMR calculation is the normalization of the formants both for sentences generated by TTF and also for those in the verification database. Basically the Lobanov method [17] was used, but the normalization was extended to the consonants v, j, l, m, n and J. The mean values and the standard deviations for F1 and F2 were determined in every speaker and in every TTF model. All further calculations are based on the normalized formant values calculated by the former means and standard deviations. The calculation of TMR values is based on the use of the correlation coefficient applied by Hermes [18], who used it for the comparison of f_0 curves in speech.

The $r(x,y)$ correlation coefficient in (1) is used to define the TMR_j^s of the compared sentences (2). The TMR_j^s is calculated separately for F1 and F2. In (2) the number of formants is expressed by j as 1=F1 or 2=F2. Two groups are expressed by the s variable, namely the group of vowels and the group of measured consonants v, j, l, m, n and J.

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

$$TMR_j^s = r\left(\hat{F}_j^{N,s}, F_j^{N,s}\right),\ j = 1,2; s = Vowel, Cons \quad (2)$$

In equation (2) $\hat{F}_j^{N,s}$ represents the normalized formant data produced by the TTF model and $F_j^{N,s}$ means the normalized formant values of the same natural sentence of the verification database after determining j and s. As a result altogether four TMR values are calculated for every compared sentence.

A wider validation has been done too, i.e. we compared the results of the TTF model with the formant measurement results of Praat [8] and Snack [6], using the default settings (Praat: time_step: 0.01 s, max num of formants: 5, max formant: 5 kHz, window length: 0.025 s, pre-emphasis: 50 Hz; Snack: lpcorder: 12, num formants: 4, sampling freq: 10 kHz, frame length: 0.01 s). Practically, automatic measurements were carried out on the sentence waveforms of the validation database, and these results were compared to the manually corrected formant values in these sentences. The result was a calculated TMR value both for Praat and Snack. These values were compared to the ones of the TTF model. The TTF models, the Praat and Snack are called together: Tools. The statistical analysis of the TMRs was performed by the Friedman and the Wilcoxon rank sum tests.

3. Results

The main verification of presented TTF models is based on the TMR values discussed in Section 2.4. Comparing the total TMR means of the different Tools, there was a significant effect of TTF models and formant trackers on levels of TMR ($p<0.001$). Means of the Tools' TMR showed decreasing order as follows: 1sp - 0.825, 5sp - 0.812, 1sp* - 0.791; Snack - 0.755, Praat - 0.527. Friedman post hoc test revealed that there was no significant difference between the 1sp* and Snack. 1sp.m* denotes the results of the five 1sp.m models, but they were compared not with themselves, but with the other 4 male speakers, which were not included in the training corpus. 1sp.f* corresponds similarly to the female speakers. 1sp* denotes all 1sp.m* and 1sp.f* models altogether.

There was a significant effect of formant on levels of TMR ($p<0.001$). F2 can be predicted better (mean: 0.867) than F1 (mean: 0.751). The gender was not significantly related to the TMR values, i.e. the male TTF models compared to the male

voice in the verification database shows 0.810, and the result for the female with the same conditions, 0.808. The mean TMR value for the vowels is 0.856, while for the consonants v, j, l, m, n and J only 0.762. The difference between these groups is significant ($p<0.001$).

Nevertheless the TMR results of the different Tools differ as regards gender, the vowel/consonant and F1/F2 group. The detailed results are shown in Figure 5 and in Table 2 for male voice and in Table 3 for female. For the comparison the following grouping was used: F1/F2; vowel/consonant; male/female.

Table 2. *TMR averages for the TTF models and also for the formant trackers. Male voice.*

TTF models vs. VDB and formant trackers vs. VDB	TMR for male speakers			
	Vowels		Consonants v, j, l, m, n, J	
	F_1	F_2	F_1	F_2
5sp.m	0.798	0.913	0.717	0.827
1sp.m	0.824	0.926	0.715	0.835
1sp.m*	0.791	0.907	0.662	0.804
Snack	0.879	0.955	0.753	0.759
Praat	0.420	0.820	0.345	0.577

Table 3. *TMR averages for the TTF models and also for the formant trackers. Female voice.*

TTF models vs. VDB and formant trackers vs. VDB	TMR for female speakers			
	Vowels		Consonants v, j, l, m, n, J	
	F_1	F_2	F_1	F_2
5sp.f	0.785	0.912	0.716	0.824
1sp.f	0.808	0.925	0.736	0.829
1sp.f*	0.774	0.906	0.691	0.793
Snack	0.602	0.821	0.609	0.662
Praat	0.547	0.641	0.453	0.417

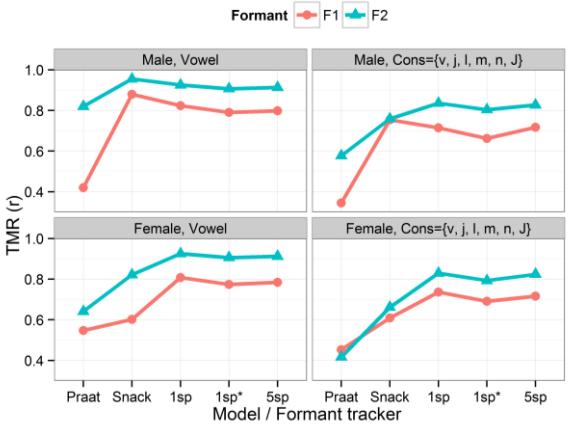

Figure 5: *The mean TMR values of the different TTF models and also of the formant trackers grouped according to gender, the vowels/consonants and F1/ F2*

The first rows show the comparison results of the 5sp models with the sentences in the verification database of their counterpart voices. These models are trained by the largest data group, and therefore we intend to use them in the final TTF converter solution. The second rows show that 1sp models perform slightly better than 5sp. However, 1sp is considered to be a baseline result of the speaker dependent model to compare the performance of the speaker independent 5sp model. (The 1sp.m row in Table 2 shows averaged results of the five individual 1sp.m models compared with their counterpart voices and sentences in VDB. Thus the training and the verification data belong to the same speaker.) In the third rows, with 1sp* models this is not the case, that is why the result is the worst among the three models.

In addition, the last two rows of both Tables show the comparison result with the two popular formant tracker algorithms mentioned before. In total, the results from Snack are slightly weaker than 5sp, and Praat is the weakest.

However, the vowel columns show that for vowels the Snack is slightly better than the 5sp.m model, but not like 5sp.f. Praat results are the lowest in every case. Concerning the columns for consonants, Snack is better for F1 than 5sp.m, but in F2 is weaker. For female consonants Snack is weaker than 5sp.f model.

In summary, the TTF converter using 5sp models gives very good predictions for F1 and F2 formant trajectories from text input. The gender can be used as an input parameter.

4. Conclusions

The presented indirect TTF model produces formant movements from text in general for Hungarian. Thus the tendencies of formant movements can be studied easily without direct measurements. The validation showed that the trajectory data produced from the parametric TTF model are not worse than those of direct formant trackers; however the model has many advantages as follows. The generated formant pattern represents indirectly the articulation motions during speaking. By the model the characteristic formant trajectories can be demonstrated easily for the language. The TTF model can be used for research, education and development. The model opens a novel way for formant prediction: easy to use and has reliable results. Mass formant prediction can be done directly from text. Language specific calculations can be performed on formant trajectories. Connecting with ASR, new ways of processing may be developed.

The FDB formant database itself can be used for studying the limit of power of the coarticulation processes during speaking. Moreover, using FDB one may increase the control of HMM-based speech synthesis similarly to [19, 20] or to create data-driven formant synthesis [9, 21]. In addition, the FDB can be used for determining and increasing the accuracy of formant measuring algorithms. The most important fact is that FDB and TTF can give a long term support for speech research in many ways. The method can be adapted to other languages as well. Live demo: http://hungarianspeech.tmit.bme.hu/ttf

5. Acknowledgements

The authors would like to thank the support of Swiss National Science Foundation via their joint research project (SCOPES Scheme) SP2: SCOPES project on speech prosody. Bálint Pál Tóth assisted in fundamental research in the frame of TÁMOP 4.2.4. A/1-11-1-2012-0001 National Excellence Program – Elaborating and operating an inland student and researcher personal support system, was realized with personal support. The project was subsidized by the European Union and co-financed by the European Social Fund.

6. References

[1] L. Deng, C. Xiaodong, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing," *Proceedings of the ICASSP* pp. 369-372, 2006.

[2] I. Jemaa, O. Rekhis, K. Ouni, and Y. Laprie, "An Evaluation of Formant Tracking methods on an Arabic Database," *Proceedings of Interspeech* pp. 1677-1670, 2009.

[3] G. Olaszy, Zs. Rácz, and K. Abari, "A formant trajectory database of Hungarian vowels," *The Phonetician* 97/98., pp. 6-13, 2008/2011.

[4] K. Abari, G. Olaszy, and Zs. Rácz, "Formant maps in Hungarian vowels – online data inventory for research, and education," *Proceedings of Interspeech*, pp. 1262-1265, 2011.

[5] G. Olaszy and K. Abari, "Az artikulációs mozgások akusztikai vetületeinek adatbázisa magyar beszédre," [Database of acoustical imprints of articulatory movements for Hungarian] (in Hungarian). *Beszédkutatás* [Speech Research] pp. 223-233, 2015.

[6] "The Snack Sound Toolkit [Computer program], Version 2.2.10." http://www.speech.kth.se/snack/, accessed Feb 15, 2015.

[7] D. Talkin, "Speech Formant Trajectory Estimation Using Dynamic Programming with Modulated Transition Costs," *JASA, S1*, p. S55, 1987.

[8] P. Boersma and Weenink, D., "Praat: doing phonetics by computer [Computer program], Version 5.4.06." http://www.praat.org/, accessed Feb 15, 2015.

[9] H. Hu, *"Towards an Improved Model of Dynamics for Speech Recognition and Synthesis"*, PhD thesis, University of Birmingham, UK, http://etheses.bham.ac.uk/3704/1/Hu12PhD.pdf, accessed Jun 8, 2015.

[10] G. Olaszy, "Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai," [Development and services of a Hungarian precisely labelled and segmented, parallel speech database] (in Hungarian). *Beszédkutatás* [Speech Research], pp. 261–270, 2013.

[11] K. Vicsi and A. Víg, "Az első magyar nyelvű beszédadatbázis," [The first Hungarian speech database] (in Hungarian), *Beszédkutatás*, pp. 163–177, 1998.

[12] K. Abari, *"A formánsmozgások statisztikai vizsgálata és modellezése a magyar magánhangzókban,"* [Modelling and analyzing the formant movements in Hungarian vowels] (in Hungarian), PhD thesis, University of Debrecen, Hungary, 2013.

[13] K. Abari and G. Olaszy, "Interaktív formánsmódosító," [Interactive tool for formant value correction] (in Hungarian) Proc. of VIII. Magyar Számítógépes *Nyelvészeti Konferencia, MSZNY,* pp. 309-315, 2011.

[14] Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, and A. W. Black, "The HMM-based speech synthesis system version 2.0," *Proc. ISCA SSW6*, pp. 294–299, 2007.

[15] A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *Journal of the Acoustical Society of America*, vol. 79, pp. 1086–1100, 1986.

[16] B. Tóth and G. Németh, "Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis," *Acta Cybernetica*, vol. 19, no. 4, pp. 715–731, 2010.

[17] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers." *Journal of the Acoustical Society of America*, (49): pp. 606–608, 1971.

[18] D. J. Hermes, "Measuring the Perceptual Similarity of pitch contours" *J. Speech, Language, and Hearing Research* vol. 41: pp. 73–82, 1998.

[19] M. Lei, J. Yamagishi, K. Richmond, Z. Ling, S. King, L. Dai, "Formant-controlled HMM-based Speech Synthesis," *Proceedings of Interspeech*, pp. 2777–2780, 2011.

[20] M.-Q. Cai, Z. Ling, L. Dai, "Formant-Controlled Speech Synthesis Using Hidden Trajectory Model," *Proceedings of Interspeech*, pp.1529-1533, 2014.

[21] G. K. Anumanchipalli, Y.-C. Cheng, J. Fernandez, X. Huang, Q. Mao, A. W. Black, "KLATTSTAT: Knowledge-based Parametric Speech Synthesis," *Proceedings of SSW* 7, pp. 206-210, 2010.