DOI: 10.2478/linpo-2020-0008

Phonetic convergence in the shadowing for natural and synthesized speech in Polish

Karolina Jankowska[1], **Tomasz Kuczmarski**[1] **& Grażyna Demenko**[1]

[1]Adam Mickiewicz University, Poznań
karolina.pieniowska@amu.edu.pl, faqster@gmail.com, lin@amu.edu.pl

Abstract: Karolina Jankowska, Tomasz Kuczmarski & Grażyna Demenko, *Phonetic convergence in the shadowing for natural and synthesized speech in Polish*. The Poznań Society for the Advancement of Arts and Sciences, PL ISSN 0079-4740, pp. 7-17

The matter of shadowing natural speech has been discussed in many studies and papers. However, there is very little knowledge of human phonetical convergence to synthesized speech. To find out more about this issue an experiment in the Polish language was conducted. Two types of stimuli were used – natural speech and synthesised speech. Five sets of sentences with various phonetic phenomena in Polish were prepared. A group of twenty persons were recorded which gave the total number of 100 samples for each phenomenon. The summary of results shows convergence in both natural and synthesised speech in set number 1, 2, 4 while in group 3 and 5 the convergence was not observed. The baseline production shown that the great majority of participants prefer *ɛn/ɛm* version of phonetic feature which was reflected in 83 out of 100 sentences. In the shadowing natural speech participants changed *ɛn/ɛm* to *ɛw/ɛ̃* in 26 cases and in 4 *ɛw/ɛ̃* to *ɛn/ɛm*. When shadowing synthesised speech shift from *ɛn/ɛm* to *ɛw/ɛ̃* in 18 sentences and 4 from *ɛw/ɛ̃* to *ɛn/ɛm*. The intonation convergence was also observed in the perceptual analysis, however the analysis of F0 statistics did not show statistically significant differences.

Keywords: shadowing, convergence, speech synthesis

1. Introduction

Spoken language is one of the most natural forms of communication for human beings. This fact has been used in technology domain which led to the development of spoken dialogue systems that are nowadays accessible for everyone who has access to a computer or a smartphone. The popularity of voice-enabled assistants such as Apple's Siri, Microsoft's Cortana or Google Now rises and people use those interfaces to perform tasks such as calling, managing calendar, checking the weather forecast or triggering a search through spoken commands (Lison & Meena 2014: 46-51). Furthermore, it is believed that language will be a basic element of human-machine communication and in the near future. Such interaction will be the main focus of the information society in the coming years. Advances in the technical solutions will lead to integration of spoken language processing with other branches of information technology. As the efficiency and quality of transducers that enable the acquisition of acoustic signals and their dimensions are reduced (cells, mobile devices), there will be a growing need for instant access to voice information.

A dialogue is an interactive process of information exchange, mainly verbal, conditioned by the involvement of both the sender and the recipient. The analysis of lexical, syntactic, prosodic and non-language, e.g. gestures and behaviour, interaction between people having conversation showed a convergence of behaviours between interlocutors. The phenomenon of interaction between speakers is well known in psycholinguistics, communication and broadly understood cognitive sciences. The fact that the shape of an utterance can be connected with various aspects such as physiology, speaker dialect, idiolect and conversational settings has been demonstrated in recent phonetic studies. It is proved that talkers converge to interlocutor in phonetic and acoustic aspects which is considered to be a property of natural dialogue (Lison & Meena 2014: 46-51). The fact that speakers can converge with interacting partners by subconscious manipulation of attributes such as accent, speaking rate, intensity, utterance duration and frequency of pauses has been studied and discussed in many articles. However, the topic of convergence to the synthesised speech remains unclear (Pardo 2013).

The way of human communication with computers triggered the questions on human convergence to synthesised speech with the convergence defined as an increase in segmental and suprasegmental similarity between two speakers. In order to investigate this subject, a speech shadowing experiment was conducted. It is a psycholinguistic experimental technique in which subjects repeat speech at a delay to the onset of hearing the phrase. The objective was to determine and compare the level of phonetic convergence for natural and synthesized speech. In the following chapters the participants, procedure and results are discussed.

1. Methodology

1.1. Stimuli

In order to conduct the experiment five sets of sentences with various phonetic phenomena in Polish were created. These included the following: (1) orthographic word-medial letters ę before consonants other than fricative, (2) orthographic word-medial letters *ą* before consonants other than fricative, (3) the letter ń in various positions, (4) realisation of *em(n)*, *om(n)* word-initially in loanwords, (5) combinations of letters *trz*, *strz*. Polish nasal vowels *ą* and *ę* are articulated in the asynchronic manner and they are realized as diphtongs consisting of an oral vowel */ɔ/* or */ɛ/* followed by a nasal stop */m/,* */n/*, */ɲ/*, */ŋ/* or nasalized approximant */w/* or / *J̃* / (Wagner 2015).

Table 1: The sentences and the occurring pronunciation variants

Orthographic word-medial letters ę before vowels other than fricative				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Ta służba to mordęga.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Wszędzie jest spory bałagan.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Wczoraj było jakieś święto.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Do dziś cierpią męczarnie.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Gęba sama się wykrzywia.	ę	*ɛw/ ɛ̃*	em	*ɛm*
Orthographic word-medial letters ą before vowels other than fricative				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Dął straszliwy wiatr.	ą	*ɔw/ ɔ̃*	o	*ɔ*
Z tej mąki nie upieczesz chleba.	ą	*ɔw/ ɔ̃*	on	*ɔn*
Obcy nie może tu rządzić.	ą	*ɔw/ ɔ̃*	on	*ɔn*
To nie jest zbyt rozsądne.	ą	*ɔw/ ɔ̃*	on	*ɔn*
Zagraj to teraz na trąbce.	ą	*ɔw/ ɔ̃*	om	*ɔm*
The letter ń in various positions				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Nad wejściem wisi końska podkowa.	ń	*ɲ°*	ń	*J̃*
Przyznawał się do duńskiego pochodzenia.	ń	*ɲ°*	ń	*J̃*
Był zupełnym jej przeciwieństwem.	ń	*ɲ°*	ń	*J̃*
Niańczą dwójkę swoich dzieci.	ń	*ɲ°*	ń	*J̃*
Ukończyła kurs dworskiego tańca.	ń	*ɲ°*	ń	*ȷ̃*
Realisation of *em(n)*, *om(n)* word-initially in loanwords				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Komfort onieśmielał ich coraz częściej.	om	*ɔm*	ą	*ɔw (/ɔ̃)*
Przybył właśnie pan konsul.	on	*ɔm*	ą	*ɔw (/ɔ̃)*
Nimfa przewróciła mu w głowie.	im	*im*	im	*iŋ*
Widać w tym dziele niezwykły kunszt.	un	*un*	un	*uŋ*
Powiedz, jaki to ma sens.	en	*ɛn*	ę	*ɛw (/ɛ̃)*
Combinations of letters *trz*, *strz*				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Ona ma już trzeciego męża.	trz	*tʂ*	cz	*t͡ʂ*
Potrzeba matką wynalazków.	trz	*tʂ*	cz	*t͡ʂ*
Jeszcze tam przytrzymaj!	trz	*tʂ*	cz	*t͡ʂ*
Nie może powstrzymać kaszlu.	strz	*stʂ*	szcz	*ʂt͡ʂ*
Basia jutro cię ostrzyże.	strz	*stʂ*	szcz	*ʂt͡ʂ*

Two speakers have been recorded for this study as the model, one male one female. The recordings were conducted in a professional studio with the use of the following equipment: overhead microphones: DPA 4066 omnidirectional headset microphone, stationary microphones: Neumann TLM 103 condenser, large diaphragm microphone, Cakewalk Sonar X1 LE Software and Roland Studio Capture hardware. The synthetic speech samples were generated using a revised version of the Polish HMM-based speech synthesizer which was, in turn, built using the HMM-based Speech Synthesis System (HTS) (Lewandowski & Jilka 2019) as first described in (Kuczmarski 2010: 221-228). The models were trained on a speech corpus originally designed for the Polish BOSS unit selection synthesizer (Demenko et al. 2007). A default speech analysis configuration for generation of the training data was used. This included data sampled at 48 kHz, with a 25 ms frame and a 5ms shift using the Hamming window. The resulting acoustic training data comprises of 48 Mel-generalized cepstrum coefficients (MGC), 24 band-aperiodic coefficients (BAP) along with a single log F0, as well as their first and second delta features. The Polish BOSS synthesiser is also used here as a text analysis tool for the HTS voice which does not provide a front-end of its own. The label files it produces for any given text input contain rich information about phones, syllables, word and phrase boundaries, as well as ToBI accents and special markers for intonational phrase types and boundaries (Breuer et al. 2010). These were first checked manually and adjusted to meet the needs of the current experiment. It was ensured that correct stress and accent marks as well as segmental variants were used in correct contexts. The resulting label files were then automatically converted into full-context HTS label format adding additional qualitative and quantitative information about the utterance to be synthesized using quintphones as the base unit. Speech was synthesized and after a brief auditory inspection we have decided to use the 2mix variant of the synthesizer output, which uses a second order Gaussian mixture probability density functions with diagonal covariance matrices, as the voice quality was seemingly the highest.

1.2. Participants

Participants were recruited at Adam Mickiewicz University in Poznań, Facular of Modern Languages. There were ten men and ten women aged 20-26. All of the twenty participants were native Polish speakers. Fourteen of them declared to learn at least one foreign language and six to learn more than two foreign languages. The participants' preference regarding the examined phonetic features was identified during the baseline phase and pictured in the following table.

Table 2: The phonetic features identified during the baseline phase

	ɛw/ɛ̃ vs. *ɛn/ɛm*		*ɔw/ɔ̃* vs. *ɔ/ ɔn*		*ɲ̊* vs. *j̃*		*ɔm/im/un/ ɛn* vs. *ɔw/ɔw*		*tʂ/stʂ* vs. *t͡ʂ/st͡ʂ*	
Women	4	46	22	28	47	3	39	11	49	1
Men	13	37	29	21	44	6	34	16	50	0
Summary	17	83	51	49	91	9	73	27	99	1

1.3. Experimental procedure

Speech shadowing is a psycholinguistic experimental technique in which subjects repeat speech at a delay to the onset of hearing the phrase. The time between hearing the speech and responding, is how long the brain takes to process and produce speech. For standard testing, the response time between perceiving and producing speech is estimated for 250 ms. This number can vary among subjects depending on certain conditions. However, for this study, the reaction time was not taken into account. The objective of this shadowing experiment is to determine the degree of convergence to human speech versus synthesized speech.

The procedure of the study was divided into four stages. The first one was the baseline production, during which the participants were reading sentences from a paper. The selection of the stimuli for the shadowing task, the second stage, depended on the realisation of the target features (pronunciation variants) when reading the sentences and it was the opposite version of what a participant produced naturally. Before the next step, a visual task was performed so as to weaken the mental representation of the baseline production. During the shadowing task four sets of five stimuli were played to the participants over the headphones (male and female voices, synthesised speech in random order). Right after the shadowing task, participants were asked to read sentences from the screen to record the post-production.

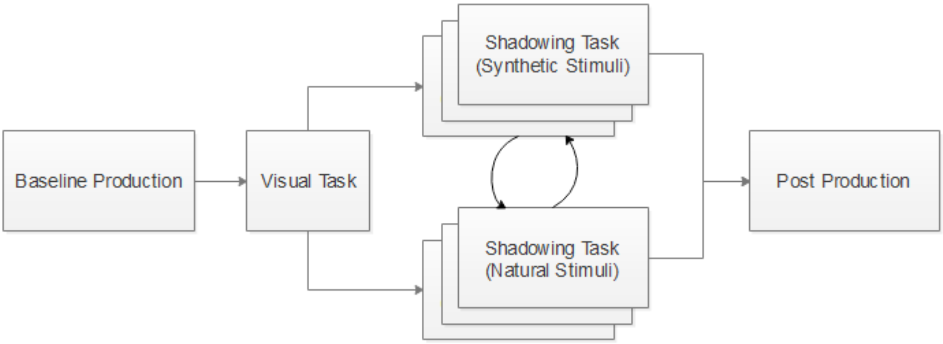

Figure 1. The diagram of experiment procedure

All the recordings were performed in a recording studio, with the same equipment as models. Each participant was asked to read the sentences casually without focusing on the meaning and expression. In the repeating task, participants were supposed to repeat the sentences not paying attention to uncertainties, manner of speech. They were assured that the whole procedure is supervised and if any problem occurs, the executor will react accordingly. Such additional instructions had to be delivered since subjects reacted inadequately to the synthesised speech, admitting that the recording is damaged. In such cases the whole procedure was repeated with another participant.

2. Results

As a result of the experiment, the spoken language corpus was created. This database can be divided into two parts – the first is model recordings, male, female and synthetic voice in two variants of pronunciation. The second part consists of 60 recordings of read speech and repeated sentences. In total, the corpus contains 66 files with a total of 1506 utterances (where utterance is understood as one sentence among those listed in the Table 1).

In order to analyse the results of the experiment three different methods were used. First is the perceptual phonetical shifts analysis, whose aim was to determine if the speaker changed the way of pronunciation of given phonetic phenomenon under the influence of a natural and synthetic stimulus. Second, the perceptual change in the realization of the phrase compared to the baseline, separately for shadowing synthesized speech and natural speech. The third was the analysis of fundamental frequency (F0) statistics.

The baseline production shown that the great majority of participants prefer *ɛn*/*ɛm* version of phonetic feature which was reflected in 83 out of 100 sentences. When shadowing natural speech, participants changed *ɛn*/*ɛm* to *ɛw* in 26 cases and in 4 *ɛw/ɛ̃* to *ɛn/ ɛm*. When shadowing synthesised speech shift from *ɛn*/*ɛm* to *ɛw* in 18 sentences and 4 from *ɛw/ɛ̃* to *ɛn*/*ɛm*. There were 12 cases of change from *ɔw/ɔ̃* to *ɔ/ɔn* and 19 from *ɔ/ɔn* to *ɔw/ɔ̃*. With the speech synthesis the numbers were 9 and 6 accordingly. The results in the third group for the natural speech was 11 *ɲ̊* to *J̃* and 8 *J̃* to *ɲ̊*, while for speech synthesis 10 and 5. For the phenomena in loanwords there were 8 switches from *ɔm/im/ un/ɛn* to *ɔ̃/ɔw/iŋ/uŋ/ɛw/ɛ̃* for natural speech and 13 for speech synthesis. Participants changed the pronunciation from *ɔ̃/ɔw/iŋ/uŋ/ɛw/ɛ̃* to *ɔm/im/un/ɛn* when shadowing natural speech in 19 cases and 7 with synthesised stimuli. In the last group the results were the same in both natural and synthetic stimuli – 11 cases for *tʂ/stʂ* to *t͡ʂ/ʃt͡ʂ* and 1 for *t͡ʂ/ʃt͡ʂ* to *tʂ/stʂ*. The following table presents the results for each phenomenon separately.

Table 3: Summary of results

Shadowing natural speech				Shadowing synthesised speech				
ɛw/ɛ̃ to *ɛn/ɛm*	*ɛn / ɛm* to *ɛw/ɛ̃*	remains *ɛw/ɛ̃*	remains *ɛn / ɛm*	*ɛw/ɛ̃* to *ɛn/ɛm*	*ɛn /ɛm* to *ɛw/ɛ̃*	remains *ɛw/ɛ̃*	remains *ɛn /ɛm*	discarted
4	26	13	57	4	18	11	63	4
ɔw/ɔ̃ to *ɔ/ɔn*	*ɔ/ɔn* to *ɔw/ɔ̃*	remains *ɔw/ɔ̃*	remains *ɔ/ɔn*	*ɔw/ɔ̃* to *ɔ/ɔn*	*ɔ/ɔn* to *ɔw/ɔ̃*	remains *ɔw/ɔ̃*	remains *ɔ/ɔn*	discarted
12	19	37	32	9	6	39	42	4
ɲ̊ to *J̃*	*J̃* to *ɲ̊*	remains *ɲ̊*	remains *J̃*	*ɲ̊* to *J̃*	*J̃* to *ɲ̊*	remains *ɲ̊*	remains *J̃*	discarted
11	8	80	1	10	5	81	1	3
ɔm/im/ un/ɛn to *ɔ̃/ ɔw/iŋ/ uŋ/ɛw*	*ɔ̃/ ɔw/iŋ/ uŋ/ɛw* to *ɔm / im / un / ɛn*	remains *ɔm/im/ un/ɛn*	remains */ɔ̃/ ɔw/iŋ/ uŋ/ɛw*	*ɔm/im/ un/ɛn* to *ɔw/ ɔw/ iŋ/uŋ/ɛw*	*/ɔ̃/ ɔw/iŋ/ uŋ/ɛw* to *ɔm/im/ un/ɛn*	remains *ɔm/im/ un/ɛn*	remains *ɔ̃/ ɔw/iŋ/ uŋ/ɛw*	discarted
8	19	65	8	13	7	58	19	3
tʂ/stʂ to t͡ʂ/ʂt͡ʂ	*t͡ʂ/ʂt͡ʂ* to *tʂ/stʂ*	remains *tʂ/stʂ*	remains *t͡ʂ/ʂt͡ʂ*	*tʂ/stʂ* to *t͡ʂ/ʂt͡ʂ*	*t͡ʂ/ʂt͡ʂ* to *tʂ/stʂ*	remains *tʂ/stʂ*	remains *t͡ʂ/ʂt͡ʂ*	discarted
11	1	88	0	11	1	88	0	0

The convergence has been observed for both, the natural and synthesised speech. In most cases the convergence to natural speech exceeded speech synthesis. The was several cases when participants did not respond to the stimuli at all or produced not whole phrase due to a problem with hearing and understanding the phrase correctly. Also, several times subjects skipped the word containing a specific phonetic feature with another which also made the phrase redundant. Such cases were considered discarded and occurred only with the speech synthesis. The degree of convergence for *t͡ʂ/ʂt͡ʂ* and *tʂ/stʂ* was equal for both types of stimuli. Similarly, the shift from *ɛw/ɛ̃* to *ɛn/ɛm.* For all the other cases the number of shifts was higher for natural stimuli with one exception *ɔm/ im/un/ɛn* to *ɔw/ɔw/iŋ/uŋ/ɛw/ɛ̃.* One possible explanation for this is that the perceptibility of specific phonetic phenomenon could be understood as a pronunciation error and the subject produced the phrase in a different manner intentionally. The following chart presents the summary of the degree of convergence for natural and synthesised speech.

The perceptual analysis shows convergence in intonation when shadowing natural and synthesised speech. The evaluators reported that in most cases the accent and intonation has been reproduced exactly the same way as the natural stimuli. For the task of shadowing synthesised stimuli, the speech was received and reported as flat.

The analysis of F0 considered global parameters of the recordings. The tool that was used for the speech features extraction and evaluation was My-Voice Analysis Python library developed by MySolutions Lab in Japan. The library is available on GitHub and is available free of charge to any person obtaining a copy of this software and associated documentation files. The tool breaks utterances and detects syllable boundaries, fundamental frequency contours, and formants without the need of manual transcription. Its

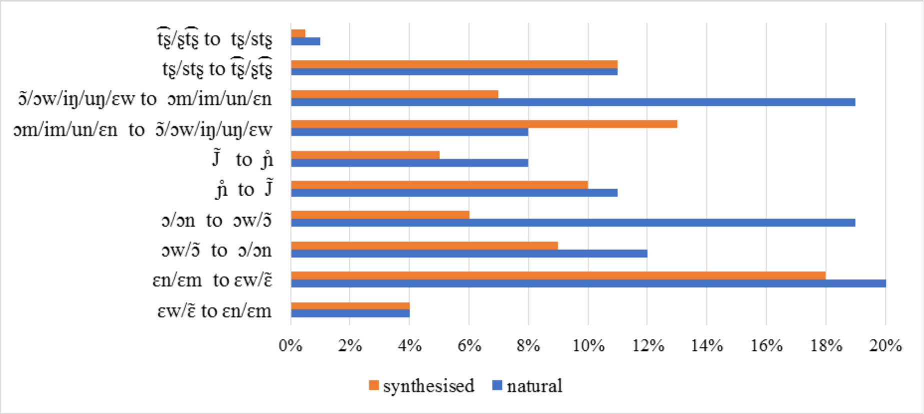

Figure 2. Summary of the degree of convergence for natural and synthesised speech

built-in functions recognise gender of the speaker, speech mood and measures pronunciation posterior score, articulation-rate, speech rate, filler words, f0 statistics. Peaks in intensity (dB) that are preceded and followed by dips in intensity are considered as potential syllable cores (Sabahi 2019). The parameters extracted with the tool were the following:

- Measure fundamental frequency distribution mean
- Measure fundamental frequency distribution SD
- Measure fundamental frequency distribution median
- Measure fundamental frequency distribution minimum
- Measure fundamental frequency distribution maximum
- Measure 25th quantile fundamental frequency distribution
- Measure 75th quantile fundamental frequency distribution

These parameters were extracted for the model stimuli and for each separate baseline, shadowing synthesised speech and shadowing natural speech. The table below presents a summary of F0 measurements for each stimulus separately.

Table 4: F0 statistics of stimuli

	Synthetic	Female voice	Male voice
f0_mean	111,78	204,55	103,53
f0_std	18,425	47,56	21,805
f0_median	113	193,75	101,2
f0_min	79,5	81	79
f0_max	180	396,5	393,5
f0_quantile25	96	170	92
f0_quan75	125	231	110,5

For each recording, similar analyses were performed, including baseline, synthetic and natural speech shadowing. All measurements were averaged for all subjects, taking into account the division into women and men. No statistically significant changes were observed under the influence of any stimulus. The average F0 measurements for natural male and female speech were very close to the standard, which makes it difficult to draw conclusions. In order to investigate this subject, a similar analysis should be performed for individual recordings separately. However, changes in the maximum value of F0 in the case of shadowing synthethic speech have been observed.

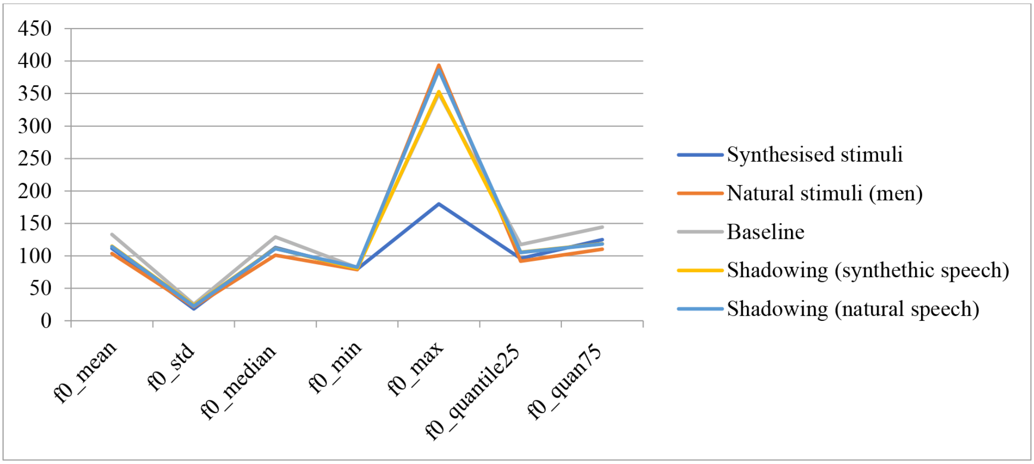

Figure 3. Average measurement results for men

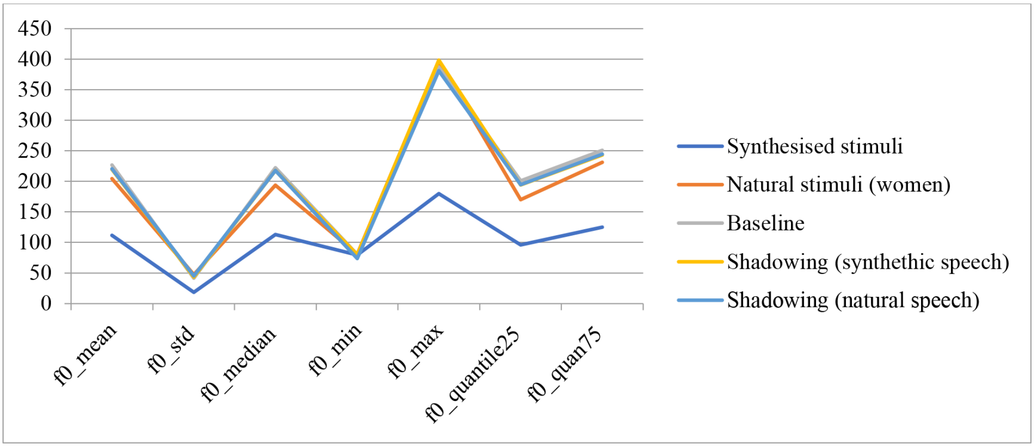

Figure 4. Average measurement results for women

3. Conclusions

The main conclusion is that the phonetic convergence occurred when shadowing both natural and speech synthesis. It can be summarised that humans do converge phonetically when interacting with synthesized speech. The degree of convergence depends on the type of the target feature (pronunciation variants). It was observed that the level of convergence of natural speech was higher than of speech synthesis.

The overall outcome of the experiment met all the expectations, allowing to draw useful conclusions from it. However, there is still a vital necessity for further work to be carried out in order to expand the data resources. It is recommended to conduct further analysis of the speech corpus, taking into account the location of given phonetic phenomena. In addition, the region of origin of the speakers was not taken into account during the study, which could have affected the differences in pronunciation. In different Polish dialects there are different tendencies that could affect the results of the experiment. Furthermore, an analysis of response time to the natural versus synthesised stimuli would be useful and interesting. In order to develop research on the convergence of people to synthesized speech, more extensive research should be carried out with more participants and with larger linguistic material.

4. Acknowledgement

This experiment has been conducted within the *Automatic analysis of phonetic convergence in speech technology systems* project no. 2014/14/M/HS2/00631 founded by National Science Centre Poland.

References

Breuer, Stefan & Stober, Karlheinz & Wagner, Petra & Abresch, Julia. 2000. *Dokumentation zum Bonn Open Synthesis System BOSS II, Unveroffentliches Dokument, IKP.* http://www.ikp.uni-bonn.de/. (Accessed 2010--09-19.)

Demenko, Grażyna & Wypych, Mikołaj & Baranowska, Emilia. 2003. Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis. *Speech and Language Technology* 7. 79-97.

Demenko, Grażyna & Klessa, Katarzyna & Szymański, Marcin & Bachan, Jolanta. 2007. The design of Polish speech corpora for speech synthesis in BOSS system (Paper presented at the conference of XII Sympozjum „Podstawowe Problemy Energoelektroniki, Elektromechaniki i Mechatroniki" PPEEm, Wisła 2007).

Gessinger, Iona & Raveh, Eran & Le Maguer, Sébastien & Möbius, Bernd & Steiner, Ingmar. 2017. Shadowing Synthesized Speech – Segmental Analysis of Phonetic Convergence (Paper presented at the conference of 18th Annual Conference of the International Speech Communication Association Stockholm, August 20-24, 2017).

Jassem, Wiktor. 2003. Polish. *Journal of the International Phonetic Association*. 103-107.

Kuczmarski, Tomasz. 2010. HMM-based Speech Synthesis Applied to Polish. *Speech and Language Technology*. Ed. Demenko, Grażyna & Wagner, Agnieszka. Poznań: Polish Phonetic Association, 2009/2010. 221-228.

Lison, Pierre & Meena Raveesh. 2014. Spoken dialogue systems: the new frontier in human-computer interaction. *Crossroads, The ACM Magazine for Students*. 46-51.

Nowakowski, Paweł & Wiatrowski, Przemysław. 2013. Informacje fonetyczno-ortograficzne w podręczniku Kultura języka polskiego. Wymowa, ortografia, interpunkcja. *Slavia Occidentalis*. 87-100.

Pardo, Jennifer S. 2013. Phonetic convergence in shadowed speech: A comparison of perceptual and acoustic measures (Paper presented at the conference of 14th Annual Conference of the International Speech Communication Association Lyon, August 25-29, 2013).

Rojczyk, Arkadiusz. 2013. Phonetic imitation of L2 vowels in a rapid shadowing task. Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference. 66-76.

Sabahi, Shahab. My Voice Analysis. March 2019. https://github.com/Shahabks/my-voice-analysis (Accessed 2019-12-03.)

Wagner, Agnieszka. 2015. Description of vowels in general and of Polish vowels in detail. https://agnieszka-wagner.weebly.com/uploads/1/5/4/8/15489492/vowels2015.pdf (Accessed 2020-01-31.)

Zen, Heiga & Nose, Takashi & Yamagishi, Junichi & Sako, Shinji & Masuko, Takashi & Black, Alan & Tokuda, Keiichi. 2007. The HMM-based speech synthesis system (HTS) version 2.0. *Proceedings of 6th ISCA Workshop on Speech Synthesis (SSW-6)*. 294-299.