Modeling Natural Communication and a Multichannel Resource: The Deceleration Effect

Andrej A. Kibrik[a,b], Grigory B. Dobrov[c] and Nikolay A. Korotaev[d]

[a] *Institute of Linguistics RAS, B. Kislovskij per. 1, Moscow, 125009, Russia*
[b] *Lomonosov Moscow State University, Leninskie gory, Moscow, 119991, Russia*
[c] *Consultant Plus, Kržižanovskogo 6, Moscow, 117292, Russia*
[d] *Russian State University for the Humanities, Miusskaya pl. 6, Moscow, 125993, Russia*

Abstract

Many AI systems imitate human communication. Specific solutions are often based on implicit theories about communication. We propose that, in order to improve performance, it is useful to consult linguistic resources registering actual communicative behavior. The study is based on a multichannel resource named RUPEX. A variety of parameters of communicative behaviors are annotated in RUPEX and can be used to improve AI systems, such as conversational agents. We focus on the deceleration effect, characteristic of elementary chunks of human speech. Specific data on deceleration can be derived from the RUPEX annotation. An assessment of deceleration in the speech produced by conversational agents is presented. Features found in linguistic annotation may provide the algorithm with certain hints on what to attend to. Annotated linguistic resources provide more direct information on what to imitate, and taking them into account may lead to better pattern recognition and therefore better speech production. RUPEX is an example of a rich resource that can help to synthesize more natural behavior.

Keywords

Conversational agents, speech, velocity, deceleration

1. Introduction

According to a popular definition, AI imitates human behavior. Among the kinds of human behavior, communication with other individuals is one of the most common activities. In accordance with that, modeling human behavior addresses various communicative processes. The domains of AI associated with modeling communication include human-computer interaction systems, social intelligence, facial recognition, speech production, etc.

How do innovative solutions developed in AI to model communication-related processes come about? It seems that engineers often rely on intuitive, implicit and ad hoc theories about human communication. For example, messenger applications create new environments for conversation. If the architecture of a messenger presupposes that turns in conversation are strictly sequential, a consequence is that it often becomes difficult to understand about a current turn which previous turn it is related to. To take another example, some anti-plagiarism systems presuppose that all words are equal; as a result, formulaic collocations such as "Relevance of the dissertation's topic is determined by the following factors" are wrongly identified as instances of plagiarism.

We propose that some of such problems may be avoided if developers of communicative systems consult what is known about communication in the science of language and, specifically, existing linguistic resources that register actual human behavior. In this paper we address the work of conversational agents that imitate human speech.

Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence, November 12-14, 2020, Moscow, Russia
EMAIL: aakibrik@gmail.com; wslcdg@gmail.com; n_korotaev@hotmail.com
ORCID: 0000-0002-3541-7637; 0000-0002-0934-3072; 0000-0002-2184-6959

Modern conversational agents work quite well, but speech is not fully natural. That may be partly due to the fact that some of the parameters, already explored by linguists and annotated in linguistic resources, are not taken into account. It may be useful to pay attention to extra parameters explored by linguists and to use resources annotated for such parameters.

In this paper we present a resource richly annotating various kinds of communicative behavior and suggest that using this resource may improve the naturalness of speech. We concentrate on one particular aspect, namely velocity of speech, but briefly mention several other parameters as well.

The paper is structured as follows. In Section 2 we review the current work on modern Text-To-Speech technology, underlying conversational agents. Section 3 presents the corpus we use to evaluate conversational agents and introduce the annotated parameters that are potentially useful for speech synthesis. In section 4 we report the results of our analysis: the evidence on velocity and deceleration in human speakers, as well as comparable features found in synthesized speech. Sections 5 and 6 contain discussion and conclusions, respectively. In the Appendix we provide illustrations of the used data.

2. Related work

Text-To-Speech technologies have been actively developing during the recent years. Particularly, speech systems based on neural networks (Neural Text-To-Speech, NTTS) have emerged as a state-of-the-art standard. Contrary to more traditional TTS systems, sequence-to-sequence NTTS frameworks do not presuppose manually annotated and complicated linguistic and syntactic features for training the model. Instead, NTTS systems use speech samples and corresponding transcripts to learn pronunciation, prosody, pauses and style. This allows one to obtain higher scores in two most frequently used metrics for evaluating the naturalness of generated spoken texts, MOS (Mean Opinion Score; see [1]) and MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor; see [2]). For example, an NTTS model reported in [3] achieved a MOS value statistically equivalent to that obtained for professional reading; while Amazon Alexa achieved a MUSHRA score of 98% for generated speech when trained on 74 speakers of 17 languages [4].

Still, several important issues remain when using the NTTS approach. First of all, high quality performance is usually achieved when training models with a large number of parameters which requires lots of resources. As shown in [5], even though the NTTS system works with a high quality for English, there remains much to improve with respect to other languages such as Japanese. The authors suggest that if a sequence-to-sequence system uses no complex linguistic features, it needs a compensation with the increased model size. Therefore, NTTS and its training data should be carefully selected to learn complex features and this limits the technology scalability. Next, NTTS models often disregard variation in prosody and intonation. Popular methods of training NTTS models such as variational encoders tend to over-smooth prosody to average values since they use Gaussian distribution as the base model. This results in all generated data being close to an average; see [6] for a critique of this approach. Moreover, NTTS systems rarely seek to reproduce speech disfluencies (filled pauses, self-repairs, etc.) inherent to natural spoken discourse, see [7]. Given these considerations, a hybrid approach has been proposed, according to which NTTS systems with additional features are used. [8] investigated a sequence-to-sequence NTTS for Japanese with additional features such as pauses, devoicing and pronunciation marks. The authors suggest that usage of such features can help predict pauses and pronunciation of devoiced vowels to increase the naturalness of the synthesized speech.

All in all, NTTS is the today's standard for industrial systems used in conversational agents. The best known examples are Amazon Polly [9], Google Cloud Speech API [10], and Yandex.SpeechKit [11]. These systems use the latest technologies in speech generation to produce speech with human-like features. In 2016 Google presented WaveNet, one of the first successful deep neural network for generating raw audio; in 2017 it was implemented in Google Assistant [12]. Yandex.SpeechKit is also based on a deep neural network. In our study, based on a Russian corpus, we decided to test Yandex.SpeechKit (Section 4.3).

3. Data and methods

This study is based on the so-called multichannel resource named Russian Pear Chats and Stories (RUPEX) [13]. The basic idea of the multichannel approach is that humans communicate not just via words but via a variety of other channels, including prosody, gesture, eye gaze, etc. We use the term "multichannel" as a replacement of the more common "multimodal" because just two modalities are involved: auditory (or vocal from the perspective of the addresser) and visual (or kinetic). Each of the modalities involves a number of channels, see Figure 1.

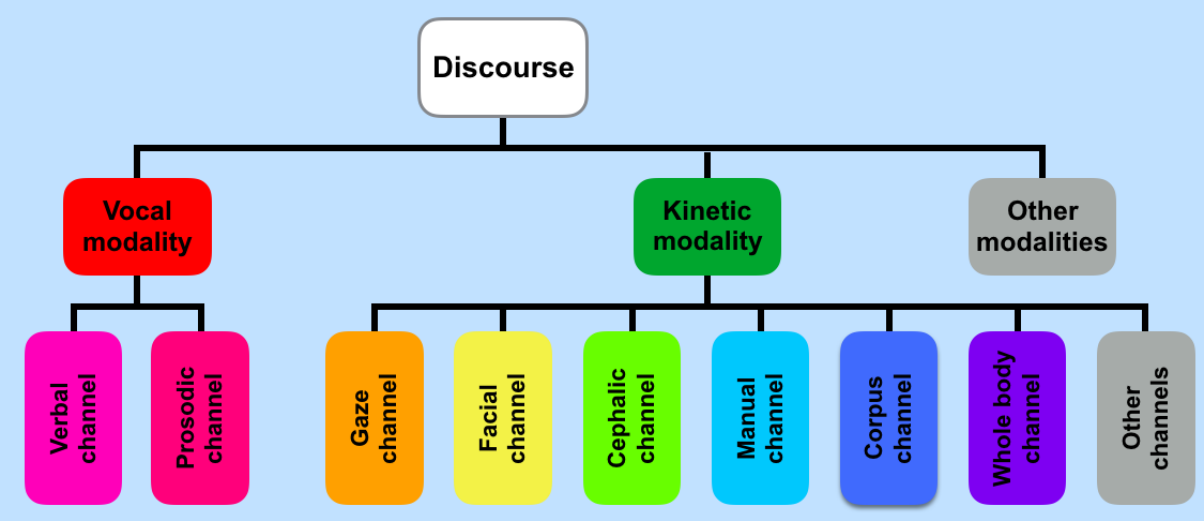

Figure 1. A scheme of multichannel discourse (from [13])

The RUPEX corpus was intended as a resource approaching in its richness natural face-to-face communication. The design of the corpus is described in [14]. The corpus consists of separate sessions, organized in a parallel way. Each session involves monologic stages during which only one person speaks, as well as a conversational stage during which three participants freely interact with each other. A session is represented in the corpus by a set of media files (audio and video recordings and eye tracker files) and a set of annotations, including vocal annotation, oculomotor annotation, manual behavior annotation, etc. The recordings and the corresponding annotations are fine-grained and register multiple phenomena, including very short ones, as well as numerous aspects thereof. See details at [15, 16]. The rich and highly detailed annotations underlie the search system implemented in the corpus, see [17, 18]

The annotation in RUPEX contains a wealth of information that can potentially be used in the synthesis of natural communicative behavior. The following tentative list of such usable parameters include five items, the first three of which are related to speech per se and the other two pertain to other channels: the channel of manual gestures and the channel of eye gaze. It must be noted that the RUPEX annotation is not limited to these parameters and includes much more diverse information. So the list below is just a sample.

- Velocity of speech
- Placement and duration of pauses
- Intonational realization of illocutionary and phasal [19: ch. 3] meanings
- Parameters of manual gestures, such as frequency of gestures, their amplitude, handedness, etc.
- Parameters of oculomotor behavior, such as direction of gaze, fixation frequency and duration.

All these kinds of data, first, are crucial for the naturalness of behavior, second, are already available in our annotation (mean and modal values can be identified) and, third, can be easily and effectively used in modeling communicative behavior. In this paper we focus on the first of these parameters, namely the velocity of speech. Some additional comments are made in the Conclusion (Section 6).

Velocity of speech (speech rate, speech tempo) has numerous functions and instantiations in human talk [20, 21, 22, 23]. [19: ch. 4] explored semantically loaded velocity that is selectively used on particular elements of speech, for example to convey meanings such as 'far vs. close to a landmark' or 'fast vs. slow process'. There is also general velocity characterizing speech of various people. There is substantial variation in the velocity of speech both across individuals and within the speech of one person. One usually differentiates between the velocity of speech production (VoSP) and the velocity of vocalization (VoV) (alternative terms: speaking rate and articulation rate [24]). VoSP is a measurement of how many units (phonemes, words, etc.) are produced per a unit of time. In contrast, VoV only accounts for periods of time when a speaker is not silent. In other words, time in VoSP includes all silent pauses, while in VoV silent pauses are excluded. This difference is important for two reasons. First, silent pauses take a substantial share of time of our speech. Second, people vary a lot in how frequent and how long their pauses are. So both measurements are relevant.

Furthermore, there is an important velocity-related feature of speech: deceleration towards the end of certain linguistic units, sometimes dubbed "drawl". In various studies deceleration was noted at the end of dialogic turns [25, 26, 27, 28] and syntactic constituents [29, 30, 31]. From our perspective, particularly important is the deceleration at the end of elementary discourse units (EDUs). EDUs are fundamental building blocks of spoken discourse; they are minimal behavioral acts that push discourse forward. A variety of terms have been used for what we call EDUs, such as syntagms, intonation units, prosodic groups, etc. Some of the studies mentioning deceleration at the end of EDUs or comparable units include [32, 33, 34, 35, 23: 177-183, 36]. The deceleration effect is robust, and listeners are very much tuned to expecting deceleration as a default. Therefore, this effect is a necessary sub-conscious parameter in assessing naturalness of speech and it must be implemented in conversational agents.

Below we explore the deceleration effect found in the speakers annotated in RUPEX and compare that with the behavior of one of the advanced conversational agents (the Yandex.SpeechKit). So we concentrate on one of the prosodic features registered in our multichannel annotation.

4. Results

In this section we report the results of our analysis: the evidence on velocity and deceleration in human speakers, as well as comparable features found in synthesized speech.

4.1. Velocity

We explored velocity features of six speakers involved in three sessions of RUPEX (session IDs 04, 22, and 23). In each session we used two speakers who produced monologic discourse: the Narrator (N) and the Reteller (R). Two types of measurements, VoSP and VoV (see Section 3), are shown in Table 1. In the case of VoV, not taking silent pauses into account, the values are naturally higher. We have used five different unit types in which velocity can be measured. All data were automatically exported from the already-existing corpus annotations. Word boundaries (and, therefore, durations) were initially identified using an algorithm developed by STEL [37] and manually checked thereafter. For the sake of simplicity, letters in transcripts stood for phonemes, vowels for syllables.

Table 1
Mean velocity of speech in six human speakers, unit/s

	Velocity of speech production					Velocity of vocalization (words + filled pauses)				
	Phonemes	Syllables	Words	Words/FPs	EDUs	Phonemes	Syllables	Words	Words/FPs	EDUs
min–max	8.61–12.03	3.65–5.22	1.92–2.56	2.14–2.68	0.57–0.69	11.54–15.98	4.90–6.91	2.56–3.36	2.79–3.5	0.73–0.89
mean	10.20	4.40	2.18	2.38	0.62	13.06	5.63	2.79	3.05	0.80

The information in Table 1 can be used as a benchmark when creating ecologically valid conversational agents (Section 4.3), as well in the subsequent discussion of the deceleration effect. In what follows, we use syllables as basic velocity units, which accords with a long-established tradition beginning from [20]. Still, it is useful to take into account other velocity measurements included in Table 1 as well; see [38] for an overview on speech rate measurements.

4.2. Deceleration

We investigated the deceleration effect in the six speakers introduced in the previous subsection. For each EDU we compared the duration of the two initial syllables (2InDur) and the two final syllables (2FinDur). Two kinds of EDUs were excluded from consideration: those containing less than four syllables, and those demonstrating a pause or another clear evidence of a prosodic boundary after the first syllable. In all, 810 EDUs were analysed, and for each one we identified the **deceleration coefficient (DC)**:

$$DC = 2FinDur / 2InDur \quad (1)$$

The data on the deceleration effect in our speakers are presented in Table 2.

Table 2
Deceleration effect in individual speakers and across all speakers

Speaker	EDUs analysed	EDUs with DC > 1	DC: mean	DC: st. dev.	DC: median
04N	148	124 (84%)	1.86	0.99	1.72
04R	150	109 (72%)	1.50	0.79	1.32
22N	108	75 (69%)	1.59	0.82	1.47
22R	129	102 (79%)	1.70	0.77	1.55
23N	109	79 (72%)	1.37	0.68	1.28
23R	166	127 (77%)	1.59	0.89	1.42
Total	810	616 (76%)	1.61	0.85	1.46

The deceleration effect in each of the speakers and in total is significant (t-test, p=0.01). Our data therefore confirm the robust character of the effect.

4.3. Comparison with Yandex.SpeechKit

We have compared the temporal features of speech in human speakers and the conversational agent Yandex.SpeechKit [11].

This utility can read text and allows for setting parameters of generated speech. It also offers a choice of several “voices”. For Russian, voices Jane and Alena are recommended. Voice Jane was available when we were at a preliminary stage of this study (April 2020). As of January 2021, voice Alena became available, and it represents a newer technology. According to the developers, Alena processes context more fully and better reproduces the details of human voice; see [39].

We used two fragments from the monologues 04N and 23N. These particular speakers were chosen because they demonstrate the highest (04N) and the lowest (23N) values of DC, see Table 2. Those fragments were selected that contained minimal instances of speech disfluencies, such as self-corrections, hesitation pauses, etc. Fragment 04N contains 73 EDUs (including 51 that were taken in consideration, see beginning of section 4.2), and fragment 23N contains 74 EDUs (including 55 taken in consideration). Transcripts of the selected fragments were rewritten in standard orthography and punctuation. Examples of our working system of spoken discourse transcription (see [36]) and of the rewritten standard representation are found in the Appendix.

Each of the rewritten fragments was read by two voices: Jane and Alena. The default (1.0) velocity value and the “Without emotion (neutral)” tone of voice were used. After speech was generated, it was

divided into EDUs in accordance with the same procedures as those used in our study of natural speech. This speech was also analysed for velocity and deceleration, just like the speech of human speakers. The data on velocity appear in Table 3.

Table 3
Velocity of speech in conversational agents and in human speakers, unit/s

Speaker	Fragment	Velocity of speech production				Velocity of vocalization (words)			
		Phonemes	Syllables	Words	EDUs	Phonemes	Syllables	Words	EDUs
Jane	04N	13.33	5.76	2.53	0.72	14.82	6.40	2.81	0.80
	23N	12.78	5.47	2.54	0.74	14.48	6.21	2.88	0.83
	Total	13.04	5.61	2.54	0.73	14.65	6.30	2.85	0.82
Alena	04N	13.00	5.62	2.46	0.73	15.45	6.67	2.93	0.87
	23N	12.37	5.30	2.46	0.71	15.24	6.53	3.03	0.88
	Total	12.67	5.45	2.46	0.72	15.34	6.60	2.98	0.87
human	04N	10.22	4.37	2.00	0.60	13.35	5.71	2.61	0.78
	23N	10.75	4.38	2.14	0.62	14.25	6.07	2.84	0.83
	Total	10.48	4.47	2.07	0.61	13.79	5.88	2.72	0.80

In contrast to Table 1, in Table 3 filled pauses are not included as conversational agents do not use them. It is interesting to note that conversational agents have velocity features higher than human speakers. In particular, if one compares the data in Tables 1 and 3, it is obvious that both Jane and Alena have VoSP values (in phonemes, syllables, and EDUs; boxes shaded in Table 3) beyond the range found in human speakers.

As for the deceleration effect, Table 4 contains the data on individual speakers. In the case of conversational agents, only the EDUs coinciding in Jane and Alena were included.

Table 4
Deceleration effect in conversational agents and in human speakers

Speaker	Fragment	EDUs analysed	DC: mean	DC: median
Jane	04N	53	1.29	1.20
	23N	55	1.49	1.38
	Total	108	1.39	1.30
Alena	04N	53	1.43	1.34
	23N	55	1.60	1.44
	Total	108	1.52	1.41
human	04N	51	1.87	1.71
	23N	55	1.38	1.28
	Total	106	1.61	1.42

In these data, just the Jane's reading of 04N is beyond the human range, as found in Table 1 (boxes shaded in Table 4). Alena decelerates stronger than Jane (significant difference according to the paired t-test, $p=0.01$). This observation accords with the perceptual feeling (and the market promotion) of Alena as a more advanced voice. The overall generalization is that both Jane's and Alena's deceleration patterns are around the bottom border of the human speakers' range, although Alena fares somewhat better. We believe that a conversational agent would benefit in naturalness if its deceleration pattern were better coordinated with what is found in the populations of speakers, as represented in annotated corpora.

5. Discussion

Our original hypothesis that insufficient naturalness of the conversational agents' speech is partly due to insufficient deceleration was based on the Jane voice only and on a much smaller dataset. (We had only included those EDUs whose two or three initial syllables coincided with full words.) When the Alena voice was added and the dataset was expanded, our hypothesis was not fully confirmed. The conversational agents mostly do fit within the lower part of the human deceleration range, and Alena as a more advanced voice decelerates better than the more basic Jane. At the same time, we have found that our expectation would be much better confirmed, if not for participant 23N who turned out unusual. We have a hypothesis on a confounding factor found in this speaker and potentially responsible for the results. Compared to five other speakers, she produces much more numerous EDU-internal silent pauses. While the nature of these pauses remains to be explored, their frequency may affect the deceleration measurements of the kind we employed in this study.

All in all, we suggest that linguistic analysis of the kind proposed in this study may be useful for the development of conversational agents. First, if engineers were consciously aware, from the very beginning, of EDUs as chunks of spoken discourse and of the deceleration effect, the training process could have been less costly. Second, there are additional parameters potentially interacting with deceleration. When listening to the speech by Jane and Alena we have noticed certain intonational oddities, in particular unnatural pronunciation of short EDUs frequently appearing in informal speech. For example, that concerns regulatory EDUs such as *vot* 'that's that, okay', short postpositional elaborations pronounced as independent sentences, etc. We are planning to test this preliminary perceptual observation by using methods of instrumental assessment.

6. Conclusion

In this paper we have discussed the deceleration effect characteristic of natural speech and asked the question whether this effect is imitated in the speech generated by Russian conversational agents. In Section 3 we also mentioned a number of other parameters of communicative behavior, annotated in RUPEX and worth testing against the actual performance of conversational agents. Some of those parameters, just like the velocity of speech phenomena, belong to the realm of the vocal modality, e.g. the placement and duration of pauses and the intonational realization of illocutionary and phasal meanings [40, 41]. A few comments are in order regarding other phenomena associated with the kinetic modality.

[14] introduced the notion of gesticulation portrait; it is a systematic individual profile of a participant's gestural behavior, usable both in the course of annotation and in post-annotation generalizations. For example, the authors proposed to formulate the participant's profile along the parameters such as (dis)inclination to stillness, (dis)inclination to using adaptors, typical gesture amplitude and typical gesture velocity. Drawing on this knowledge, informed annotation decisions can be made. By analogy with the present study, one may suggest that embodied conversational agents imitating natural human behavior must have gesture features comparable to what is found in the annotated corpus.

As for the oculomotor channel, there are patterns of eye gaze revealed with the help of eye trackers. For instance, [42] demonstrated that at the monologic stages of the RUPEX sessions participants tend to have long fixations on their interlocutors alternating with brief fixations on the environment, while the relative frequency of these two kinds of gaze targets differs for speakers and listeners. See [43] for an implementation of eye gaze patterns in robot-to-human interaction. Also cf. studies attempting to apply the knowledge of how eye gaze operates in communication to AI systems, such as [44, 45].

Generally, we suggest that some patterns of natural communicative behavior belonging to various communication channels may be too complicated to be recognized from raw data. If one uses annotated data when developing an algorithm, solutions may be less costly. Features found in linguistic annotation may provide the algorithm with certain hints on what to attend to. For example, the algorithm that is aware of EDUs may discover the deceleration effect more easily. Annotated linguistic resources provide more direct information on what to imitate, and taking them into account may lead to better pattern

recognition and therefore better speech production. RUPEX is an example of a rich resource that can help to synthesize more natural behavior.

Acknowledgements

We are grateful to Anastasia Khvatalina who performed the annotation that we used in calculating the deceleration coefficient in human speakers. Research underlying this study was supported by the Russian Foundation for Basic Research, project 19-012-00626.

References

[1] R. Dall, J. Yamagishi, and S. King, Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation, in: Proc. 7th International Conference on Speech Prosody 2014, 2014, pp. 1012–1016. doi: 10.21437/SpeechProsody.2014-191.

[2] J. Latorre et al., Effect of Data Reduction on Sequence-to-sequence Neural TTS, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7075–7079. doi: 10.1109/ICASSP.2019.8682168.

[3] J. Shen et al., Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779–4783. doi: 10.1109/ICASSP.2018.8461368.

[4] J. Lorenzo-Trueba et al., Towards Achieving Robust Universal Neural Vocoding, in: Interspeech 2019, 2019, pp. 181–185. doi: 10.21437/Interspeech.2019-1424.

[5] Y. Yasuda, X. Wang, and J. Yamagishi, Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis, arXiv:2005.10390 [cs, eess, stat], 2020. doi: arxiv-2005.10390.

[6] Z. Hodari, O. Watts, and S. King, Using generative modelling to produce varied intonation for speech synthesis, in: 10th ISCA Speech Synthesis Workshop, 2019, pp. 239–244. doi: 10.21437/SSW.2019-43.

[7] R. Dall, M. Tomalin, and M. Wester, Synthesising Filled Pauses: Representation and Datamixing, in: Proc. 9th ISCA Speech Synthesis Workshop, 2016, pp. 7–13. doi: 10.21437/SSW.2016-2.

[8] T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis, in: 10th ISCA Speech Synthesis Workshop, 2019, pp. 166–171. doi: 10.21437/SSW.2019-30.

[9] Amazon Polly, Amazon Web Services, Inc. URL: https://aws.amazon.com/polly (accessed Jan. 17, 2021).

[10] Speech-to-Text: Automatic Speech Recognition, Google Cloud. URL: https://cloud.google.com/speech-to-text (accessed Jan. 17, 2021).

[11] Yandex.Cloud – Yandex SpeechKit. URL: https://cloud.yandex.ru/docs/speechkit/ (accessed Jan. 17, 2021).

[12] WaveNet launches in the Google Assistant, Deepmind, Oct. 4 2017. URL: https://deepmind.com/blog/article/wavenet-launches-google-assistant (accessed Jan. 17, 2021).

[13] Russian Multichannel Discourse. URL: https://multidiscourse.ru/main/?en=1 (accessed Jan. 17, 2021).

[14] A. A. Kibrik and O. V. Fedorova, A «Portrait» Approach to Multichannel Discourse, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2019, pp. 1908–1912. URL: https://www.aclweb.org/anthology/L18-1300.

[15] Russian Multichannel Discourse – Corpus. URL: https://multidiscourse.ru/corpus/?en=1 (accessed Jan. 17, 2021).

[16] Russian Multichannel Discourse – Principles of annotation. URL: https://multidiscourse.ru/annotation/?en=1 (accessed Jan. 17, 2021).

[17] Search in RUPEX. URL: https://multidiscourse.ru/search/?locale=en#!/query (accessed Jan. 17, 2021).

[18] N. A. Korotaev, G. B. Dobrov, and A. N. Khitrov, RUPEX Search: Online Tool for Analyzing Multichannel Discourse, in: B. M. Velichkovsky, P. M.Balaban, V. L. Ushakov (eds.), Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics - Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, October 11-16, 2020, Moscow, Russia, Springer Nature, to appear.
[19] S. V. Kodzasov, Issledovanija v oblasti russkoj prosodii [Studies in Russian Prosody], Jazyki Slavjanskix Kul'tur, Moscow, 2009.
[20] F. Goldman-Eisler, Psycholinguistics: Experiments in spontaneous speech, 1st edition, Academic Press, London, 1968.
[21] L. Ceplītis, Analiz rečevoj intonacii [Analysis of speech intonation], Zinātne, Riga, 1974.
[22] J. Laver, Principles of Phonetics, Cambridge University Press, Cambridge, 1994.
[23] O. F. Krivnova, Ritmizacija i intonacionnoe členenie teksta v 'processe reči-mysli': opyt teoretiko-èksperimental'nogo issledovanija [Rythmization and intonation articulation in the 'Speech-Thought process'], Doctoral Thesis, Moscow State University, Moscow, 2007.
[24] M. P. Robb, M. A. Maclagan, and Y. Chen, Speaking rates of American and New Zealand varieties of English, Clinical Linguistics & Phonetics 18-1 (2004), 1–15. doi: 10.1080/0269920031000105336.
[25] S. Duncan, On the structure of speaker–auditor interaction during speaking turns, Language in Society 3-2 (1974), 161–180. doi: 10.1017/S0047404500004322.
[26] J. Local and G. Walker, How phonetic features project more talk, Journal of the International Phonetic Association 42-3 (2012), 255–280. doi: 10.1017/S0025100312000187.
[27] S. Bögels and F. Torreira, Listeners use intonational phrase boundaries to project turn ends in spoken interaction, Journal of Phonetics 52 (2015), 46–57. doi:10.1016/j.wocn.2015.04.004.
[28] C. Rühlemann and S. Th. Gries, Speakers advance-project turn completion by slowing down: A multifactorial corpus analysis, Journal of Phonetics 80 (2020), 100976. doi: 10.1016/j.wocn.2020.100976.
[29] D. H. Klatt, Vowel lengthening is syntactically determined in a connected discourse, Journal of Phonetics 3-3 (1975), 129–140. doi: 10.1016/S0095-4470(19)31360-9.
[30] T. Cambier-Langeveld, The Domain of Final Lengthening in the Production of Dutch, Linguistics in the Netherlands 14-1 (1997), 13–24. doi: 10.1075/avt.14.04cam.
[31] A. E. Turk and S. Shattuck-Hufnagel, Multiple targets of phrase-final lengthening in American English words, Journal of Phonetics 35-4 (2007), 445–472. doi: 10.1016/j.wocn.2006.12.001.
[32] L. V. Bondarko, Zvukovoj stroj sovremennogo russkogo jazyka [Sound system of modern Russian], Prosveščenie, Moscow, 1977.
[33] A. Cruttenden, Intonation, Cambridge University Press, Cambridge ; New York, 1986.
[34] W. J. M. Levelt, Speaking: From intention to articulation, The MIT Press, Cambridge, MA, 1989.
[35] W. Chafe, Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing, The University of Chicago Press, Chicago, 1994.
[36] A. A. Kibrik, N. A. Korotaev, and V. I. Podlesskaya, Russian spoken discourse: Local structure and prosody, in: S. Izre'el, H. Mello, A. Panunzi, T. Raso (eds.), In search of basic units of spoken language: A corpus-driven approach, volume 94 of Studies in Corpus Linguistics, John Benjamins, Amsterdam, 2020, pp. 37–76. doi: 10.1075/scl.94.01kib.
[37] STEL – Speech technologies. URL: http://speech.stel.ru/ (accessed Jan. 17, 2021).
[38] T. S. Kendall, Speech Rate, Pause, and Linguistic Variation: an Examination through the Sociolinguistic Archive and Analysis Project, Dissertation, Duke University, 2009.
[39] Yandex.Cloud Documentation – Yandex SpeechKit – Speech synthesis. URL: https://cloud.yandex.ru /docs/speechkit/tts/ (accessed Jan. 17, 2021).
[40] N. A. Korotaev, Pauzy xezitacii v rasskaze i razgovore: sopostavitel'nyj količestvennyj analiz [Hesitation pauses in narratives and conversations: A quantitative comparison], in: Proceedings of the international conference "Corpus Linguistics-2019", St. Petersburg, 2019, pp. 48–54.
[41] N. A. Korotaev, Reč' i žestikuljacija v dialoge vs. monologe: opyt kontroliruemogo sopostavlenija [Speech and gesticulation in dialogues vs. monologues: Controlled comparison], in: "Word and Gesture" (Grishina's readings), Moscow, 2020, pp. 14–16.
[42] O. V. Fedorova, Zritel'noe vnimanie govorjaščego i slušajuščego na monologičeskix ètapax estestvennoj kommunikacii: razvivaja idei A. Kendona [Visual attention of the speaker and listener

at the monological stages of natural communication: Developing Kendon's ideas], Socio- and psycholinguistic studies 8 (2020), 17–25.
[43] A. A. Zinina, N. A. Arinkin, L. Ja. Zaydelman, and A. A. Kotov, The role of oriented gestures during robot's communication to a human, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 18 (2019), 800–808.
[44] R. Ishii, Y. I. Nakano, and T. Nishida, Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze, ACM Trans. Interact. Intell. Syst., 3-2 (2013), 11:1–11:25. doi: 10.1145/2499474.2499480.
[45] C. Liu, Y. Chen, M. Liu, and B. E. Shi, Using Eye Gaze to Enhance Generalization of Imitation Networks to Unseen Environments, IEEE Transactions on Neural Networks and Learning Systems (2020), 1–1. doi: 10.1109/TNNLS.2020.2996386.

Appendix

Table 5 presents an initial excerpt from the original (produced by a human speaker) 04N fragment discussed in Section 4.3. For details on the transcription system, see [36]. The transcript was created in the Russian Cyrillic writing system, but here we provide its transliteration as well as an English translation for each EDU. Also, EDUs' deceleration coefficients are given in a separate column.

Table 5
An excerpt from the 04N fragment: Transcript and deceleration coefficient (DC) values

Line ID	Transcription	DC
N-vE063	\Pacan edet na /velosipede, Some fella is riding a bicycle,	0.59
N-vE064	velosiped /bol'šoj, the bicycle is big,	3.58
N-vE065	kakie \starye /sovetskie, like old Soviet [ones],	0.77
N-vE066	esli ty /videla, if you have seen [those],	3.27
N-vE067	(ə 0.14) s <k=> || očen' vysokoj /ramoj mužskoj, with a very high men's [bicycle] frame,	3.27
pN-022	(1.04)	
N-vE068	(ɐɯ 0.37) velosiped mal'ciku /velik, the bicycle is too big for the boy,	0.57
N-vN019	(ɥ 0.30)	
N-vE069	on edet ne ʷv /sedle, he is riding not [sitting] on the seat,	1.57
N-vE070	a-a (ɥ 0.33) nu \tak, but like,	n/a
pN-023	(0.58)	
N-vE071	kak \stoja krutit \↑pedaliʷ, like pedalling in the standing position,	1.09
pN-024	(0.42)	
N-vN020	(ɥ 0.38)	
N-vE072	/proezžaet, passes by,	3.05
N-vE073	vidit –↑gruši ʷ, sees the pears,	2.55
N-vE074	–↑ostanavlivaetsjaʷ, stops,	2.54
pN-025	(0.76)	

Line ID	Transcription	DC
N-vE075	(ɯ 0.66) \velosiped ne stavit na /podnožku, doesn't prop the bicycle on the kickstand,	1.89
N-vE076	kladët ego na –↑zemlju, lays it on the ground,	1.75
N-vN021	(ɥ 0.31)	
N-vE077	smotrit vniz /naverx, looks down and up,	1.25
pN-026	(0.51)	
N-vE078	smotrit čto fermerᵊ == sees that the farmer…	1.59
N-vE079	vidit \fermer sobiraet tam –gruši, sees that the farmer is collecting pears there,	1.08
N-vE080	polnost'ju pogloščën ètim /zanjatiem, completely absorbed by this task,	0.88
N-vE081	ničego ne \↑zamečaet, notices nothing,	2.38
N-vN022	(ɥ 0.50)	
N-vE082	snačala mal'čik xočdet vzjat' /odnu ↑grušu, at first the boy wants to take only one pear,	2.54
N-vN023	(ɥ 0.49)	
N-vE083	/potom-m ponumaet čto-o (ʾ 0.45) ničto emu ne /grozit, then he understands that he runs no risk,	0.73
N-vE084	dovol'no bespalevno berët celuju –↑korzinu, and easy as pie he takes a whole basket,	3.91
pN-027	(0.20)	
N-vE085	(s bol'šim \–trudom,) with a big effort,	1.71
N-vN024	(ɥ 0.43)	
N-vE086	on eë \vzgromozdil (ʾ 0.21) na /velosiped, he loaded it on the bicycle,	2.68
pN-028	(0.89)	
N-vE087	(ʾ 0.17) i tak i p= || \uexal. and left like that.	1.74

Below follows the rewritten version of the full 04N fragment that was read by two voices (Jane and Alena) available at Yandex.SpeechKit [11]. For the sake of reproducibility, the text is provded in the standard Russian orthography. Also, we provide a free English translation of the whole fragment.

Пацан едет на велосипеде, велосипед большой, какие старые советские, если ты видела, с очень высокой рамой мужской. Велосипед мальчику велик, он едет не в седле, а ну так, - как стоя крутит педали, проезжает, видит груши, останавливается, велосипед не ставит на подножку, кладёт его на землю. Смотрит вниз наверх, видит фермер собирает там груши, полностью поглощён этим занятием, ничего не замечает. Сначала мальчик хочет взять одну грушу, потом понимает что ничто ему не грозит, довольно беспалевно берёт целую корзину (с большим трудом), он её взгромоздил на велосипед и так и уехал. С корзиной. Едет-едет, с этой тяжеленной корзиной, по дороге, такая эта дорога каменистая, неровная, едет ему навстречу девочка. Постарше, вот такая с косами. Они разминаются на дороге, разминулись, мальчик на нее оглядывается, у него слетает шляпа (ветром её видимо сносит), и велосипед наталкивается на большой камень посреди дороги. Мальчик падает, груши рассыпаются, он лежит, велосипедом его придавило, он садится, потирает коленку, ногу потирает, ушибся. У него такие

высокие носки, он один из них приспускает, чтоб посмотреть на свою ногу, оглядывается: стоят, смотрят на него ещё три пацана, разнокалиберные. Один из них самый крупный, выше всех, видимо какой-то там главарь. Самый мелкий, у него такая игрушка, похоже как будто теннисная ракетка, к ней шарик - (привязан), и шарик стучит о деревяшку, и он постоянно (как очень нервный) ей стучит, на протяжении всего своего присутствия в кадре.

A guy is riding a bicycle, a big bicycle, like old Soviet ones, as you may have seen, with a very high men's bicycle frame. The bicycle is too big for the boy, and he's not sitting on the seat but is rather like pedaling in an upright position. He passes by, then sees the pears, stops, and doesn't prop the bicycle on the kickstand but lays it on the ground. He looks down and up and sees that the farmer is picking up pears and, completely absorbed by this task, doesn't notice anything. At first, the boy wants to take just one pear, but then he realizes that he runs no risk here and he impudently takes a whole basket (with a big effort), loads it on the bicycle, and leaves like that, with this basket. He rides and rides, with this heavy basket, along the road, and the road is rocky and rough, and a girl rides towards him. She is a bit older, with plaits. They pass by each other on the road, they passed each other, the boy looks back at her, his hat flies off his head (apparently blown with the wind), and the bicycle runs over a big rock in the middle of the road. The boy falls down, the pears get scattered all around, he is lying under the weight of the bicycle, then he sits up, rubs his knee, rubs his leg, he is hurt. He has these high socks, he lowers one of them a little bit to inspect his leg, he looks around and there are three other guys standing there and looking at him, such a motley crew. One of them is the bulkiest, he's taller than the others, he seems some kind of a leader. The smallest guy, he has this toy, like a tennis racket, a ball is attached to it, this ball is hitting the paddle, and this guy is constantly (as if being very nervous) playing with the racket, for the whole duration of this scene.