Studying Mutual Phonetic Influence With a Web-Based Spoken Dialogue System*

Eran Raveh[1,2] [0000−0003−4411−9663], Ingmar Steiner[1−3] [0000−0001−6415−5915], Iona Gessinger[1,2] [0000−0001−5333−9794], and Bernd Möbius[1] [0000−0003−3065−9984]

[1] Language Science & Technology, Saarland University, Germany
[2] Multimodal Computing and Interaction, Saarland University, Germany
[3] German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany
raveh@coli.uni-saarland.de

Abstract. This paper presents a study on mutual speech variation influences in a human-computer setting. The study highlights behavioral patterns in data collected as part of a shadowing experiment, and is performed using a novel end-to-end platform for studying phonetic variation in dialogue. It includes a spoken dialogue system capable of detecting and tracking the state of phonetic features in the user's speech and adapting accordingly. It provides visual and numeric representations of the changes in real time, offering a high degree of customization, and can be used for simulating or reproducing speech variation scenarios. The replicated experiment presented in this paper along with the analysis of the relationship between the human and non-human interlocutors lays the groundwork for a spoken dialogue system with personalized speaking style, which we expect will improve the naturalness and efficiency of human-computer interaction.

Keywords: Spoken dialogue systems · Phonetic convergence · Human-computer interfaces.

1 Introduction

With expanding research on, and growing use of, spoken dialogue systems (SDSs), a main challenge in the development of human-computer interaction (HCI) systems of this kind is making them as close as possible to human-human interaction (HHI) in terms of naturalness, fluency, and efficiency. One aspect of such HHIs is the relationship of influences between the interlocutors. Influence here means changes in one interlocutor's conversational behavior triggered by the behavior of the other interlocutor. We refer to changes that make the interlocutors' behaviors more similar as *convergence*. Convergence can occur on the level of different modalities and with respect to various aspects of the conversation, like eye gaze, gestures, lexical choices, speech, and more. In this paper, we concentrate on influences on the phonetic level, i.e., *phonetic convergence*. More specifically, we examine variations in pronunciation. As speech is the principal modality used for communicating with SDSs, we believe it is an especially important modality to study in the field of HCI. Simulating and triggering convergence

* Funded by the German Research Foundation (DFG) under grants STE 2363/1 and MO 597/6.

on the phonetic level, as found in HHI, may contribute a lot to the naturalness of dialogues of humans with computers. SDSs with such personalized speech style may offer more natural and efficient interactions, and move one more step away from the *interface metaphor* [5] toward the *human metaphor* [3].

The novel system introduced in Section 3, which will be made freely available, tracks the states of segment-level phonetic features during the dialogue. All of the analyses are automated and run in real time, which not only saves time and manual work typically needed in convergence studies, but also makes the system more suitable for integration into other applications. In Section 4, we use this newly introduced system and recordings collected as part of a shadowing experiment to examine the relationship of mutual influences between a (simulated) user and the system. Using these signals, the system provides both visual and numerical evidence of the mutual influence between the interlocutors over the course of the interaction.

2 Background and Related Work

Integrating support for changes in the speech signal into computer systems may enhance HCI and provide improved tools for studying convergence in HCI. [17] discusses the advantages of systems that dynamically adapt their speech output to that of the user, and the challenges involved in developing and using these systems.

2.1 Phonetic Convergence

According to [18], phonetic convergence is defined as an *increase in segmental and suprasegmental similarity between two interlocutors* (e.g., [26]). In contrast to *entrainment*, we use the term *convergence* to describe dynamic, mutual, and non-imposing changes, with potentially non-polarized realizations. Phonetic convergence has been found to different degrees in conversational settings [13]. There is evidence for phonetic convergence being both an internal mechanism [20] and socially motivated [9]. Previous studies of phonetic convergence in spontaneous dyadic conversations have focused on speech rate [25] and timing-related phenomena [22], pitch [8], intensity [12], and perceived attractiveness [15]. Phonetic convergence is often examined in the scope of shadowing experiments, in which the participants are asked to repeat utterances spoken by an interlocutor [7]. This is typically done with single words as targets. The experiment showcasing our system in Section 4 uses whole sentences as stimuli, in which the target features are embedded, making it a semi-conversational HCI setting.

2.2 Adaptive Spoken Dialogue Systems

Various studies have investigated entrainment and priming in SDSs, aiming to better understand HCI dynamics and improve task-completion performance. [14], for example, focused on dynamic entrainment and adaptation on the lexical level. Others, like [16], concentrated on word frequency. [19] examined changes in both lexical choices and word frequency. While these studies addressed the changes in experimental, scripted

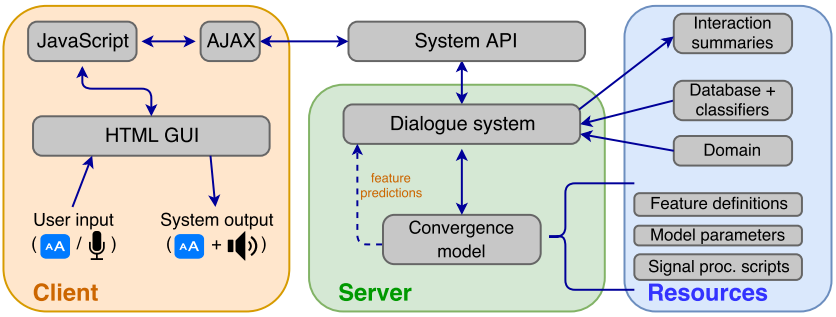

Fig. 1: An overview of the system architecture. The background colors distinguish client components, server components, and external resources that can be customized.

scenarios, the theoretical foundations for studying these changes in spontaneous dialogue exist as well [2]. [6] provide examples of online adaptation for dialogue policies and strategies.

It is important to note that while all of the studies mentioned above examine various aspects of dialogues, none of those are related to speech – the primary modality used to interact with SDSs. Studying convergence of speech in an HCI context is made possible with more natural synthesis technology, which gives fine-grained control over parameters of the system's spoken output. Many systems that deal with adaptation of speech-related features focus on prosodic characteristics like intonation or speech rate. [10] sheds light on acoustic-prosodic entrainment in both HHI and HCI via the use of interactive avatars. [1] found that users' speech rate can be manipulated using a simulated SDS. Similar results were found when intensity changes in children's interaction with synthesized text-to-speech (TTS) output were examined [4].

All of the above provide solid ground for further investigation of phonetic convergence in HCI using SDSs.

3 System

The system introduced here is an end-to-end, web-based SDS with a focus on phonetic convergence and its analysis over the course of the interaction. Besides placing convergence in the spotlight, it is designed to be flexible and to meet the researcher's needs by offering a wide range of component customizations (see Section 3.2). Its online access via a web browser makes it scalable and simple for the end-user to operate. The system's architecture and functionality are described in Section 3.1, its graphical user interface (GUI) and operation in Section 3.3, and an example of its utilization is demonstrated in Section 4. Ultimately, it offers an experimentation platform for studying phonetic convergence, with emphasis on the following:

Temporal analysis – offering real-time visualization of the interlocutors' relations with respect to selected phonetic feature over the course of the interaction.

Customizability – allowing the user to experiment with different scenarios by configuring parameters and definitions in many of the system's components.

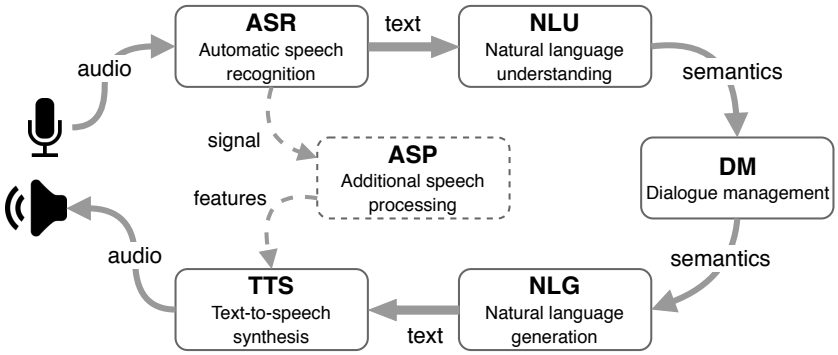

Fig. 2: The architecture of the dialogue system component. The ASP component and connections (in dashed lines) between the ASR and TTS are responsible for performing additional speech processing for phonetic convergence.

Online scalability – connecting multiple web clients to a server, allowing users to use it anywhere without installation, and helping experimenters to collect data remotely.

3.1 Architecture

As the system aims to offer a customizable playground for experimenting and studying phonetic convergence in HCI, a key aspect of its architecture is the separation between client-side, server-side, and external resources (see Figure 1). All of the resources and configuration files needed for designing the interaction are located on the server. Running the client and server on different machines allows users to interact with the system using only a web browser.

As shown in Figure 2, the **dialogue system** component consists of typical SDS modules such as natural language understanding (NLU) and a domain manager (DM), but also contains an additional speech processing (ASP) module [23]. This module is responsible for processing the audio and extracts the features required by the convergence model. While the NLU component uses merely the transcription provided by the ASR, the ASP module analyzes the speech signal itself. More specifically, it tracks occurrences of the defined features and passes the measured values to the convergence model, which, in turn, forwards the tracked feature parameters to the TTS synthesis component.

3.2 Models and Customizations

The **computational model for phonetic convergence** used in our system is described in [24]. Different behavioral patterns with respect to phonetic convergence that were observed in HHI and HCI experiments can be simulated by combinations of the model's parameters presented in Table 1. All of the parameters can be modified in the system's configuration file.

Table 1: Summary of the convergence models parameters in order of use in the convergence pipeline. Parameters with an asterisk are defined for each feature separately.

allowed range*	allowed value range for new instances
history size	maximum number of exemplars in pool
update frequency	frequency to recalculate feature's value
calculation method*	method to calculate pool value
convergence rate	weight given to pool value when recalculating
convergence limit*	the maximum degree of convergence allowed

The entire convergence process is based on the the **tracked phonetic features** that are defined as "convergeable", i.e., prone to variation, and is triggered whenever the ASR component detects a segment containing a phoneme associated with one or more of these features. Each feature is represented by a key-value map, in which the parameters from Table 1 are defined. A classifier can be associated with each feature to provide real-time predictions for both the user's and the system's realizations of that features, as demonstrated in Figure 3. With this information available, more meaningful insights can be gained into the dynamics of phonetic changes in the dialogue.

The **dialogue domain** is specified in an XML-based file. Parameters are introduced for triggering models or rules, or for processing by external modules of the SDS, as in the case of the ASP module. The format of the domain file makes it easy to define new scenarios for the system, such as a task-specific dialogue, general-purpose chat, or an experimental setup.

Speech processing is a central aspect of the system. Different models can be used, e.g., for improving performance or changing the language or the ASR module or the output voice of the TTS module.

3.3 Graphical User Interface

The system's GUI consists of three main areas:

In the **chat area**, the interaction between the user and the system is shown in a chat-like representation. Each turn's utterance appears inside a chat bubble with different colors and orientations for the user and the system. The turns are also numbered, to better track the dialogue progress and analysis shown by the plots in the graph area.

In the **interaction area**, the user can interact with the system with written or spoken input. Text-based interactions progress through the dialogue (if applicable) and trigger any subsequent domain model, but will not affect convergence-related models, since there is no audio input to process. Spoken input can be provided either by speaking into the microphone or via audio files with pre-recorded speech. The latter option is especially useful for simulating specific user input, or for reproducing a previous experiment, as done in Section 4.1.

In the **graph area**, each of the tracked features is visualized in a separate plot, and new data points are automatically added whenever a new instance of the feature is detected. Hovering over a data point in a graph reveals additional information, such as the turn in which it was added, or the realized variant of the feature produced in that

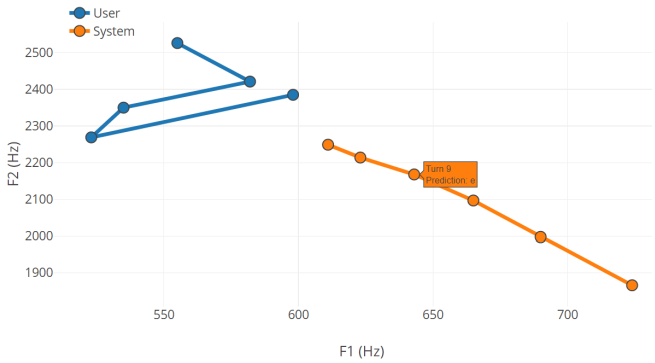

Fig. 3: A screenshot of the plot area showing the states of the feature [ɛː] vs. [eː] (in formant space) during an interaction. The system's internal convergence model (orange, bottom right) gradually adapts to the user's (blue, upper left) detected realizations. A prediction of the feature's current realization is given for both interlocutors. The annotation box marks the turn in which the system aggregated enough evidence from the user's utterances and changes its pronunciation from [ɛː] to [eː].

turn as predicted by its classifier. These dynamic, interactive plots make it possible to shed light on how the interlocutors influence each other, even if unawarely, throughout their exchanges. Figure 3 shows a graph with several accumulated data points.

4 Showcase: Examining Convergence Behaviors

We used the stimuli and recordings collected during the shadowing experiment detailed in [7] to look into types of participant convergence behaviors with respect to the features examined in the experiment (see Table 2). The experiment is designed to trigger and analyze phonetic convergence by confronting the participants with stimuli, in which certain phonetic features are realized in a way different from their own. The use of the system for this purpose results in an automated, reproducible execution, with additional insights like classification of feature realizations and dynamic visualizations in the web GUI. The classifiers were trained offline on the data points acquired from analyzing the stimuli. However, the system also supports incremental, online re-training whenever requested by the user, such as every time the convergence model is updated. A sequential minimization optimization (SMO) [21] implementation of the support vector machine (SVM) classifier was used for training. Each turn's number and prediction are added as an interactive annotation to the dynamic graph of the relevant features, as shown in Figure 3. Finally, using the system, the experiment is transformed into an automated dialogue scenario, which enhances its HCI nature.

4.1 Finding Behavioral Patterns

We simulated the shadowing experiment by configuring the domain file with a definition of the transition between the phases, as well as the flow within each phase. This

Table 2: Examples of stimuli sentences, each containing one target feature.

Sentence	Feature
War das Gerät sehr teuer? *Was the device very expensive?*	[ɛː] vs. [eː] in word-medial ⟨ä⟩
Ich bin süchtig nach Schokolade. *I am addicted to chocolate.*	[ɪç] vs. [ɪk] in word-final ⟨-ig⟩
Wir besuchen euch bald wieder. *We will visit you soon again.*	[n̩] vs. [ən] in word-final ⟨-en⟩

automates the procedure and adapts it to the participant's pace. Additional variables are defined and handled as well, allowing the system to track the experiment's flow and state. Participants were simulated by using their recorded speech from the original experiment. After each turn, all relevant SDS modules were triggered based on the simulated participant's input. The stimulus order from the original experiment was preserved. In this section, we focus on the validation for the feature [ɛː] vs. [eː] as a representative example for the phonetic adaptation capability of the system. Although the classified realization is binary ([ɛː] or [eː]), the underlying representation used by the model is gradual. Both of these views on the feature can be seen in the graph area, as shown in Figure 3.

The degree of convergence was examined per utterance in the shadowing phase of the experiment. Three main groups emerged, each with a different behavior: One group of participants showing little to no tendency to converge (changes in $\leq$10 % of their utterances), the second, with varying degrees of convergence (10 % to 90 %), and a third group of participants who were very sensitive to the stimuli's variation ($\geq$90 %). We refer to these groups as *Low*, *Mid*, and *High*, respectively. The feature's classifier was determined on the fly, so that the prediction for each utterance was made based on the stimulus type to which the participant was listening. As Table 3 shows, the *Low* and *High* groups are both of significant size, indicating that these two distinct behaviors exist in the data and can be spotted by the system.

In addition, we wanted to validate the separation between these behaviors. To this end, we regarded the shadowing phase as an annotation task, where the annotators are the predictors of the user and the system. Note that 100 % similarity would mean complete convergence to every stimulus, which cannot be reasonably expected (cf. [7]). The Cohen's kappa (κ) values[4] of the *Low* group are expected to be the lowest, as a lesser degree of convergence was found among these participants. By the same logic, the *High* group is expected to have the highest agreement, and the *Mid* to have values between the two other groups. Indeed, this hypothesis holds: weak agreement was found in the *Low* group, strong agreement in the *High* group, and a value close to 0 for the *Mid* group, indicating no consistent behavior.

[4] as calculated by the *kappa2* command of the *irr* R package (v0.84), https://cran.r-project.org/package=irr

Table 3: A summary of the measures for similarity and agreement between the predictor annotations of user and model productions in the shadowing phase.

	Similarity (%)	Agreement (κ)	Size (%)
Low	<1	−0.57 ***	23
Mid	22	−0.15 *	50
High	26	0.81 ***	27
All	48	−0.11 *	100

5 Conclusion and Future Work

We have introduced a system with an integrated spoken dialogue system (SDS), which can track and analyze mutual influence on the phonetic level during the interaction based on an internal convergence model. This combines work done in the fields of phonetic convergence and adaptive SDSs. Many aspects of the system are customizable, which makes it flexible in terms of possible supported scenarios. This includes multiple parameters defining the target phonetic features, which allows experimentation with different features. The system can also run on a separate server, which makes it easier to scale its use.

In addition, we replicated a shadowing experiment using the system, which examined phonetic convergence regarding certain features. Three main user behaviors were found with respect to their tendency to change their pronunciation based on the system's input. This sheds light on possible relations and dynamics between a user and a system in HCI. We believe that this shows that phonetic convergence can be studied using our SDS, and that this is one step forward toward personalized, phonetically aware SDSs, which will enable more natural and efficient interaction.

Future work will pursue two independent directions: regarding phonetic convergence, supporting more features will make the system more comprehensive and useful for studying a wider range of phenomena. Specifically, adding support for suprasegmental (i.e., prosodic) features will enable replication of experiments similar to e.g., [11] in the same manner as in Section 4. Regarding user acceptance, it would be interesting to examine whether users show any preference toward an SDS that converges to their speech on the phonetic level, and whether they would change their speaking style based on the system's output, forming an interaction with mutual and dynamic convergence. The first research question can be tested by comparing user interaction with a baseline system and one with convergence capabilities, and evaluating the users' performance and satisfaction. The second research question can be investigated by comparing the users' speech when interacting with either system configuration. Additionally, to test the system's influence on users' speech, the users can train with an intelligent computer-assisted language learning (CALL), such as a computer-assisted pronunciation training (CAPT) system, which will change its learner model based on their input. Task completion rate, performance accuracy, and completion time metrics can be used to evaluate how helpful the system is.

References

1. Bell, L., Gustafson, J., Heldner, M.: Prosodic adaptation in human-computer interaction. In: 15th International Congress of Phonetic Sciences (ICPhS). pp. 2453–2456. Barcelona (2003), https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_2453.html
2. Brennan, S.E.: Lexical entrainment in spontaneous dialog. In: International Symposium on Spoken Dialogue (ISSD). pp. 41–44. Philadelphia, PA, USA (1996)
3. Carlson, R., Edlund, J., Heldner, M., Hjalmarsson, A., House, D., Skantze, G.: Towards human-like behaviour in spoken dialog systems. In: Swedish Language Technology Conference (SLTC). Gothenburg, Sweden (2006)
4. Coulston, R., Oviatt, S., Darves, C.: Amplitude convergence in children's conversational speech with animated personas. In: Interspeech. pp. 2689–2692. Denver, CO, USA (2002), http://www.isca-speech.org/archive/icslp_2002/i02_2689.html
5. Edlund, J., Heldner, M., Gustafson, J.: Two faces of spoken dialogue systems. In: Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems. Pittsburgh, PA (2006)
6. Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., Young, S.: On-line policy optimisation of Bayesian spoken dialogue systems via human interaction. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8367–8371. Vancouver, BC, Canada (2013). https://doi.org/10.1109/ICASSP.2013.6639297
7. Gessinger, I., Raveh, E., Le Maguer, S., Möbius, B., Steiner, I.: Shadowing synthesized speech – segmental analysis of phonetic convergence. In: Interspeech. pp. 3797–3801. Stockholm, Sweden (2017). https://doi.org/10.21437/Interspeech.2017-1433
8. Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., Steiner, I.: Convergence of pitch accents in a shadowing task. In: Speech Prosody. Poznań, Poland (2018), in press
9. Kim, M., Horton, W.S., Bradlow, A.R.: Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. Laboratory Phonology **2**(1), 125–156 (2011). https://doi.org/10.1515/labphon.2011.004
10. Levitan, R.: Acoustic-prosodic Entrainment in Human-human and Human-computer Dialogue. Ph.D. thesis, Columbia University, New York, NY, USA (2014). https://doi.org/10.7916/D8GT5KCH
11. Levitan, R., Beňuš, Š., Gálvez, R.H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., Hirschberg, J.: Implementing acoustic-prosodic entrainment in a conversational avatar. In: Interspeech. pp. 1166–1170. San Francisco, CA, USA (2016). https://doi.org/10.21437/Interspeech.2016-985
12. Levitan, R., Hirschberg, J.: Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Interspeech. pp. 3081–3084. Florence, Italy (2011), http://www.isca-speech.org/archive/interspeech_2011/i11_3081.html
13. Lewandowski, N.: Talent in Nonnative Phonetic Convergence. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany (2012). https://doi.org/10.18419/opus-2858
14. Lopes, J., Eskenazi, M., Trancoso, I.: Automated two-way entrainment to improve spoken dialog system performance. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8372–8376. Vancouver, BC, Canada (2013). https://doi.org/10.1109/ICASSP.2013.6639298
15. Michalsky, J., Schoormann, H.: Pitch convergence as an effect of perceived attractiveness and likability. In: Interspeech. pp. 2253–2256. Stockholm, Sweden (2017). https://doi.org/10.21437/Interspeech.2017-1520

16. Nenkova, A., Gravano, A., Hirschberg, J.: High frequency word entrainment in spoken dialogue. In: ACL Human Language Technologies (HLT). pp. 169–172. Columbus, OH, USA (2008), http://aclweb.org/anthology/P08-2043
17. Oviatt, S., Darves, C., Coulston, R.: Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. ACM Transactions on Computer-Human Interaction **11**(3), 300–328 (2004). https://doi.org/10.1145/1017494.1017498
18. Pardo, J.S.: On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America **119**(4), 2382–2393 (2006). https://doi.org/10.1121/1.2178720
19. Parent, G., Eskenazi, M.: Lexical entrainment of real users in the Let's Go spoken dialog system. In: Interspeech. pp. 3018–3021. Makuhari, Chiba, Japan (2010), http://www.isca-speech.org/archive/interspeech_2010/i10_3018.html
20. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences **27**(2), 169–190 (2004). https://doi.org/10.1017/S0140525X04000056
21. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Burges, C.J.C., Schölkopf, B., Smola, A.J. (eds.) Advances in Kernel Methods, pp. 185–208. MIT Press (1999)
22. Putman, W.B., Street, R.L.: The conception and perception of noncontent speech performance: Implications for speech-accommodation theory. International Journal of the Sociology of Language **1984**(46), 97–114 (1984). https://doi.org/10.1515/ijsl.1984.46.97
23. Raveh, E., Steiner, I.: A phonetic adaptation module for spoken dialogue systems. In: Workshop on the Semantics and Pragmatics of Dialogue (SemDial). pp. 162–163. Saarbrücken, Germany (2017)
24. Raveh, E., Steiner, I., Möbius, B.: A computational model for phonetically responsive spoken dialogue systems. In: Interspeech. pp. 884–888. Stockholm, Sweden (2017). https://doi.org/10.21437/Interspeech.2017-1042
25. Schweitzer, A., Walsh, M.: Exemplar dynamics in phonetic convergence of speech rate. In: Interspeech. pp. 2100–2104. San Francisco, CA, USA (2016). https://doi.org/10.21437/Interspeech.2016-373
26. Walker, A., Campbell-Kibler, K.: Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. Frontiers in Psychology **6**(546), 1–18 (2015). https://doi.org/10.3389/fpsyg.2015.00546