TRACKING THE TEMPORAL RELATION BETWEEN SPEAKER RECOGNITION AND PROCESSING OF PHONETIC INFORMATION

Carola Schindler and Eva Reinisch

Institute of Phonetics and Speech Processing (IPS), Ludwig Maximilian University, Munich, Germany
{carola.schindler | evarei}@phonetik.uni-muenchen.de

ABSTRACT

To study the temporal relation between speaker recognition and the processing of phonetic information, we conducted a visual-world eyetracking study, in which speaker identification (two male voices) and word recognition (onset-overlapping competitors) were assessed simultaneously by presenting speaker-item combinations as the visual referents. Results showed that participants could reliably identify speakers and items in all conditions. As for the temporal uptake of information and competition between visual referents of speakers and items, we found that across conditions speaker competition was stronger than phonetic competition. Only when both speaker and item referents were ambiguous, did phonetic competition manifest itself. This suggests that speaker information is processed rapidly such that phonetic competition can be minimised. We conclude that the visual-world paradigm can be further extended to study the interaction of different types of information in the speech signal.

Keywords: speech perception, word recognition, online processing, speaker recognition, eyetracking

1. INTRODUCTION

Speech contains not only linguistically relevant phonetic and lexical information, but also information about the identity of the speaker. It has been shown that as a result of this dual function, the two types of information interact in speech perception [6-8,13]. Speaker information can thereby hinder or help word recognition. When listening to word lists spoken by multiple speakers, listeners have more difficulty in deciding whether they have heard a certain word before than when listening to only one speaker [10]. However, once listeners tune in to a speaker's specific speech characteristics or even to the accent of a speaker group, then word recognition is facilitated [16]. Crucially, the phonetic make-up of words has been shown to influence speaker identification and discrimination [2,3,17]. Certain phonetic segments such as vowels, fricatives and nasals are more indicative of speaker identity than others [2].

Although these and other types of interaction between speaker and phonetic information are well documented, most studies on the *online* processing of speech have so far focused on the uptake of phonetic information only (e.g. [1,4,18]). It has been shown that at every point during the word recognition process, listeners take into account the available phonetic information to modulate hypotheses about the words they hear (e.g., [8,14,15]). Relatively less is known about the online process of speaker recognition. The present study set out to shed light on this issue.

We asked whether the visual-world eyetracking paradigm that has been used to reveal the uptake of fine phonetic detail [15], could also provide insights into the relative timing of recognising the speaker of a given utterance (see [5] for a first attempt). Specifically, we assessed whether we could track the temporal relation of the uptake of speaker and phonetic information during word recognition. Listeners performed a visual-world eyetracking task, listening to words spoken by two speakers and viewing displays with speaker-item combinations, that is, a picture of a speaker combined with the picture of an object. Therefore, to identify the intended referent, speaker as well as phonetic information could be used (see Methods for details). In a *speaker condition*, participants had to use both speaker and phonetic information to identify the intended visual referent, because the screen contained the same item twice, but coupled with different speakers. In the *item condition*, speaker information could be used to speed up the recognition process but using phonetic information "alone" would also solve the problem.

In both conditions we expected to see competition between phonetically similar items as has been shown in previous studies using the visual-world paradigm [1]. Critically, however, the question was whether in the present study there would also be competition between the two

speaker referents, and if so, in what temporal relation this would stand relative to the uptake of phonetic information.

2. METHOD

2.1. Participants

Twenty-four students from the University of Munich participated for a small payment. They were native speakers of Standard German and between 19 and 28 years old. None of them reported any language or hearing impairments.

2.2. Materials

128 picturable German nouns with an initial CV sequence were selected such that 16 words each started with one of the consonants /p, t, b, d, m, n, f, s/. All words were produced by two young male adult native speakers of German. For each word, a picture was selected using Google picture search. In addition, two pictures of young male adults were chosen to represent the speakers.

2.3. Design

Speakers and items were combined into two types of displays. In the **speaker condition**, both speakers were assigned the same two items, for example, both speakers were displayed with *Fernrohr* "telescope" and *Filzhut* "felt hat". Items started with the same initial consonants followed by different vowels. In the **item condition** four different items were displayed. The items of one speaker started with different consonants followed by the same vowel. The other speaker was displayed with different items starting with the same consonants but the other vowel (Figure 1).

Figure 1: Example display for the item condition showing the two speakers and the items *Fernrohr, Silber, Filzhut, Sessel* "telescope, silver, felt hat, armchair".

Participants' task was to listen to the recorded words and identify the intended item and the speaker who said it. Therefore, in the speaker condition where participants saw both speakers assigned the same two items, they were forced to identify the speaker to find the target. In the item condition where four different items were on display they could identify the intended item by just using the phonetic information. When speaker 1 was the target speaker then the other picture of speaker 1 would be the speaker competitor. When *Fernrohr* was the target then either the other picture of *Fernrohr* (speaker condition) or *Filzhut* (item condition) would be the phonetic competitor.

2.4. Procedure

The Experiment consisted of three parts. In the first part, participants were familiarised with the pictures and the words they represented. That is, participants viewed all pictures with their labels written underneath. In the second part, participants learned to associate the speakers' voices with the corresponding speaker pictures. First, they were presented with each speaker producing the same eight words in random order. Words covered all possible word-initial consonants. Then participants were presented with pictures of both speakers and had to indicate which of the two had produced a given word. Words were the same as before. Feedback was given.

For the third, main part, participants were fitted with an Eyelink 1000 system (SR Research) to monitor their eye movements. On each trial participants were presented with one of the displays described above. 1200 ms later one of the words was presented auditorily over headphones. Listeners' task was to click with the computer mouse on the matching speaker-item combination. Every participant responded to 128 targets, half of which were spoken by speaker 1, half by speaker 2; half were from the speaker condition, half from the item condition. For each participant, targets were chosen such that the eight word-initial consonants were distributed evenly across speakers and conditions. For each participant the target remained unpredictable if pictures or displays were repeated to obtain answers for another item, the other condition or the other speaker. Target positions on the displays were distributed evenly. Across participants, each target was drawn equally often considering also every item-speaker-condition combination. 1000 ms after a participant responded, the next trial started. Every 10th trial a drift correction was carried out.

3. RESULTS

All participants chose the correct item in over 95% of the trials. Only correct trials were analysed. Figure 2 shows fixation proportions over time on the four speaker-item combinations. Black solid lines mark the target; black dashed lines the speaker competitor (i.e., the other item paired with the same speaker); grey solid lines the phonetic competitor (the other speaker paired with the phonetically similar or same item, depending on condition) and grey dashed lines the unrelated competitor (other speaker phonetically different item). The vertical lines represent word/consonant onset, consonant offset, vowel offset and word offset – shifted by 200 ms. 200 ms is the time that is usually assumed to elapse between an acoustic signal and eye fixations related to this signal [1]. The time of the initial consonants and vowels was normalised to allow for comparisons between the different consonants.

Table 1: Results of speaker and item condition in T1 (200-500 ms): fixation preferences between different competitors.

	Speaker Condition		Item Condition	
Comparison	*b*	*t*	*b*	*t*
speaker – unrel.	0.35	3.13	0.33	2.87
phonetic – unrel.	0.31	2.90	0.06	0.6
speaker – phonetic	0.03	0.29	0.27	2.32

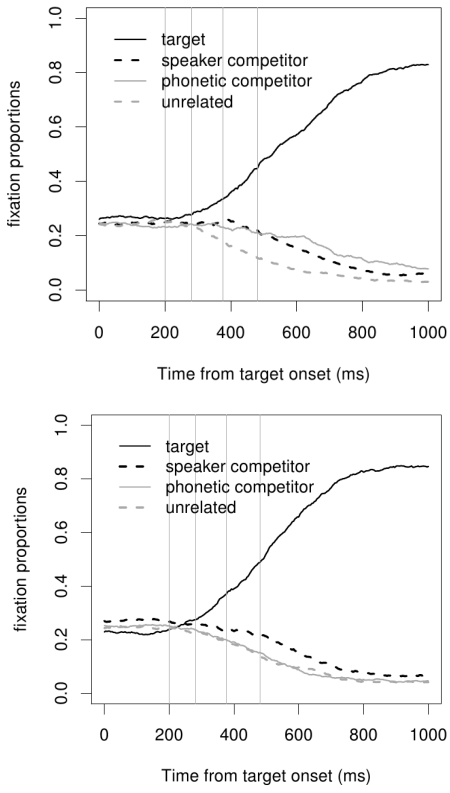
Two different time windows were analysed: T1 from 200 ms to 500 ms (see Table 1) and T2 from 500 ms to 900 ms. T1 was chosen to reflect fixations during the processing of the word (average word offset at 480 ms). T2 spanned from word offset to the point where target fixations stopped rising and competitor fixations went back to a minimum.

For each speaker-item condition, linear mixed-effects models were fitted. The dependent variables were fixation preferences between the different types of competitors. For analyses, fixation proportions were logistically transformed. As fixed factors the models only contained an intercept term. If significantly different from zero ($t > 2$) one of the competitors was fixated more than the other with the valence of the regression weight indicating the direction. Participant was entered as a random factor. Model comparisons indicated that an additional random intercept over items did not change the results.

During T1 the target was favoured quite early over all competitors in both conditions. In the speaker condition the speaker and the phonetic competitor were fixated more than the unrelated competitor and not differently from each other. That is, listeners temporarily considered the other referents that matched at least in either speaker or phonetic information (here the whole word). In the item condition the speaker competitor was favoured over both the phonetic and the unrelated competitor with the latter two not differing from each other. That is, listeners considered the item that had been paired with the target speaker despite its phonetic mismatch at word onset.

In T2, results were consistent with those of T1 with two exceptions: First, the target fixations were lower in the speaker than the item condition ($b = -0.51$; $t = -5.10$); this difference was not significant in T1; Second, in the speaker condition, the phonetic competitor was fixated more than the speaker competitor ($b = -0.30$; $t = -2.17$). That is, participants were more certain about which word has been said than who said it. These results suggest that even after word offset, participants tested whether the other picture of the item (paired with the other speaker) could have been the target.

Figure 2: Fixation proportions over time in the speaker condition (top panel) and the item condition (lower panel).



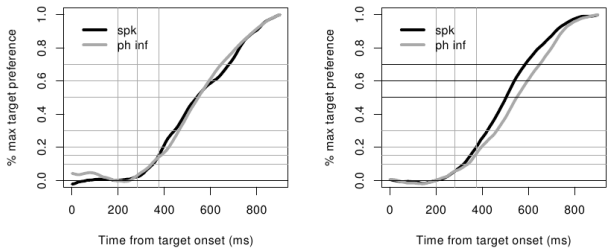
Additional analyses including place and manner of articulation of the target's initial consonants as fixed factors did not show any significant results.

That is, no differences were found for the different word-initial consonants.

In order to compare the temporal processing of phonetic and speaker information more closely, we conducted a peak-latency or maximal-effects analysis using a jackknife-based method ([9]; see [8,11] for applications to eye tracking data).

That is, time points were calculated at which participants' fixations on the target speaker (either of the two pictures) reached certain percentages of the maximum. Percentages were 10, 15, 20, 30, 50, 60 and 70% to cover a relatively large range [11]. The same was done for fixations on the target word (speaker condition) or target consonant (item condition), again pooled over the two matching options. The points in time when these percentages of the maximum effects were reached for speaker and phonetic information were then compared by *t*-tests. Distributions for these tests were based on jackknifed data [12]. Figure 3 shows the results.

Figure 3: Proportion of maximal target preference in the time window between 200 and 900 ms after target onset in the speaker condition (left panel) and the item condition (right panel).



In the speaker condition, no significant temporal difference could be found between speaker and phonetic information at any percentage point of the maximum. Note that in this condition all words started with the same consonant, hence speaker information could have been used earlier than phonetic information to recognise the target. In the item condition, the effect of speaker preceded the effect of the phonetic pattern at 50% (t = -2.56), 60% (t = -3.01) and 70% (t = -2.97) of the maximum. That is, while initially (at lower percentages) no difference could be found, later the maximum of the speaker effect was approached faster than the maximum of the phonetic effect.

4. DISCUSSION

The present study addressed whether the visual-world eyetracking paradigm that has been used to reveal the temporal uptake of fine phonetic detail [15] in online word recognition, could also provide insights into the relative timing of speaker recognition. In a task combining speaker and word recognition, participants identified the visual referents at high accuracy. The target was identified rapidly with competitors differing mainly in the rate of their decrease rather than competition in form of a rise in fixations. The speaker condition where listeners had to use speaker information in order to find the intended referent, appeared more challenging than the item condition where the target could have been found without recognising the speaker.

Given this difference in importance of using speaker information between conditions it may be surprising that (1) in the speaker condition we found speaker *and* phonetic competition with even more fixations on the phonetic competitor after word offset. (2) In the item condition where use of speaker information was "optional" we found speaker competition *but no* phonetic competition. While (1) can be explained by the difficulty of the task given that both the speakers and items were fully ambiguous, (2) may appear less straightforward. One reason for speaker rather than item competition may be that the phonetic overlap was too short to trigger phonetic competition (i.e., only the consonant overlapped). Previous research has shown that even sub-phonemic mismatches influence phonetic competition [9] and here the coarticulatory influence of mismatching vowels between phonetic competitors could have triggered such a mismatch. However, this still does not explain why the competitor that started with a *different* consonant would allow for speaker competition (i.e., the target speaker was also paired with an item starting with a different consonant than the target).

The timing of speaker and item effects suggests that speaker recognition proceeded somewhat faster than item recognition. That is, listeners appeared to first decide on the speaker and then the item. Follow-up experiments will have to clarify whether this temporal order may be strategic. Given that this was the first study to show speaker competition in a visual-world paradigm, we decided on maximising the salience of the speaker: using speaker information was crucial in the speaker condition and the speakers were displayed slightly larger than the items. Additionally the same two speakers were shown over the whole experiment while items varied on every trial.

Having shown that visual-world eyetracking can be used not only to track the uptake of phonetic (i.e., lexical) information it will be a useful tool in the future to provide several new insights on how speaker information is processed in relation to lexical access.

7. REFERENCES

[1] Allopenna, P. D., Magnuson, J. S., Tanenhaus, M. K. 1998. Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language* 38, 419-439.

[2] Amino, K., Sugawara, T. and Arai, T. 2005. The Correspondences between the Perception of the Speaker Individualities Contained in Speech Sounds and Their Acoustic Properties. *Interspeech* Lisbon, 2025-2028.

[3] Andics, A., McQueen, J. M., van Turennout, M. 2007. Phonetic Content Influences Voice Discriminability. *Proc. 16th ICPhS Saarbrücken*, 1829-1832.

[4] Cooper, R. M. 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology* 6, *84-107*.

[5] Creel, S. C., Aslin, R. N., Tanenhaus, M. K. 2008. Heeding the voice of experience: The role of talker variation in lexical access. *Cognition* 106, 633-664.

[6] Creel, S. C., Bregman, M. R. 2011. How Talker Identity Relates to Language Processing. *Language and Linguistics Compass* 5/5, 190-204.

[7] Creel, S. C., Tumlin, M. A. 2011. On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language* 65, 264-285.

[8] Cutler, A., Andics, A., Fang, Z. 2011. Inter-dependent categorization of voices and segments. In W.-S. Lee, & E. Zee (Eds.), *Proc. 17th ICPhS* Hong Kong, 552-555.

[9] Dahan, D., Tanenhaus, M. K. 2004. Continuous Mapping From Sound to Meaning in Spoken-Language Comprehension: Immediate Effects of Verb-Based Thematic Constrains. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30, 498-513.

[10] Goldinger, S. D. 1998. Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review* 105, 251-279.

[11] McMurray, B., Clayards, M. A., Tanenhaus, M. K., Aslin, R. N. 2008. Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review* 15, 1064-1071.

[12] Miller, J., Patterson, T., Ulrich, R. 1998. Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology* 35, 99-115.

[13] Nygaard, L. C., Sommers, M. S., Pisoni, D. B. 1994. Speech perception as a talker-contingent process. *Psychological Science* 5, 42-46.

[14] Reinisch, E., Sjerps, M. 2013. The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics* 41, 101-116.

[15] Salverda, A. P., Kleinschmidt, D., Tanenhaus, M. K. 2014. Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language* 71, 145-163.

[16] Samuel, A. G., Kraljic, T. 2009. Perceptual learning for speech. *Attention, Perception, & Psychophysics* 71, 1207-1218.

[17] Schindler, C., Reinisch E., Harrington, J. 2014. Perceptual speaker discrimination based on German consonants. *IAFPA* Zurich.

[18] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., Sedivy, J. C. 1995. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science* 268, 1632-1634.