Journal of Phonetics 41 (2013) 101–116

Contents lists available at SciVerse ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context

Eva Reinisch *, Matthias J. Sjerps

Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 21 February 2012
Received in revised form
20 December 2012
Accepted 5 January 2013

ABSTRACT

Speech perception is dependent on auditory information within phonemes such as spectral or temporal cues. The perception of those cues, however, is affected by auditory information in surrounding context (e.g., a fast context sentence can make a target vowel sound subjectively longer). In a two-by-two design the current experiments investigated when these different factors influence vowel perception. Dutch listeners categorized minimal word pairs such as /tɑk/–/taːk/ (''branch''–''task'') embedded in a context sentence. Critically, the Dutch /ɑ/–/aː/ contrast is cued by spectral and temporal information. We varied the second formant (F_2) frequencies and durations of the target vowels. Independently, we also varied the F_2 and duration of all segments in the context sentence. The timecourse of cue uptake on the targets was measured in a printed-word eye-tracking paradigm. Results show that the uptake of spectral cues slightly precedes the uptake of temporal cues. Furthermore, acoustic manipulations of the context sentences influenced the uptake of cues in the target vowel immediately. That is, listeners did not need additional time to integrate spectral or temporal cues of a target sound with auditory information in the context. These findings argue for an early locus of contextual influences in speech perception.

1. Introduction

When we listen to speech, our perceptual system continuously evaluates a multitude of auditory cues. These cues are used to modulate hypotheses about which words are being heard. The relative importance of different cues depends on the listeners' specific phoneme inventories. In Dutch, for instance, both duration and spectral cues are among the most important determinants of vowel identity (Adank, van Hout, & Smits, 2004; Adank, van Hout, & van de Velde, 2007). However, in order to properly evaluate these cues, listeners have to deal with a large amount of variation (Peterson & Barney, 1952). Such variation partly arises because speakers have different vocal-tract properties and can differ in speaking style or dialect. One of the important ways in which listeners deal with such variation is by making use of acoustic information that is available in the context (Johnson, 2005; Kluender, Coady, & Kiefte, 2003; Nearey, 1989). For example, listeners take into account that, in fast speech, all segments shorten (e.g., Crystal & House, 1982, 1988). That is, following a context that is spoken quickly, a sound is subjectively perceived to be longer than it actually is. Critically, this influence can result in a change of the perceived identity of a speech sound (see e.g., Miller, 1981, 1987 for overviews). This shows that cues to phoneme identity are perceived relative to their context rather than only on the basis of absolute acoustic values.

It remains unclear, however, at what stage in the cognitive architecture of speech perception speech cues are influenced by auditory properties of their context. On the one hand, speech cues might first be evaluated locally (i.e., at the phoneme or syllable level), and only be integrated with more distal context at a later stage in perception (e.g., at or after the initial stages of lexical access). On the other hand, contextual influences could express themselves immediately. That is, as the acoustic signal unfolds sound-internal cues are always immediately interpreted relative to acoustic context. A third alternative is that the locus of contextual influences differs depending on the speech cue that is involved. For example, for spectral cues, integration could be immediate whereas for durational cues integration could occur at a later stage, or vice versa. To evaluate these accounts, the present study focused on Dutch listeners' perception of the /ɑ/–/aː/ vowel contrast as occurring in Dutch minimal word pairs (e.g., /tɑk/–/taːk/, ''branch''–''task''). Importantly, Dutch listeners rely on both duration and spectral cues to identify these vowels

* Corresponding author. Now at: Ludwig Maximilian University Munich, Department of Phonetics and Speech Processing, Schellingstraße 3, 80799 Munich, Germany. Tel.: +49 15787610098.
E-mail address: evarei@phonetik.uni-muenchen.de (E. Reinisch).

0095-4470/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2013.01.002

(Escudero, Benders, & Lipski, 2009; Gerrits, 2001; Nooteboom & Coden, 1984; van Heuven, van Houten, & De Vries, 1986). First, we created a vowel that is perceived to be ambiguous between /ɑ/ and /aː/. In the first parts of our two experiments this ambiguous vowel was further manipulated (i.e., disambiguated) by increasing or decreasing its second formant (F_2) frequency and by lengthening or shortening its duration. The resulting vowels were presented in a neutral (i.e., unmanipulated) context sentence. In the second parts of the experiments, the target vowel was kept in a perceptually ambiguous range of durational and spectral values but the words were placed into a context sentence that was manipulated. This was again done by increasing or decreasing the F_2 in all vowels and by lengthening and shortening the duration of all segments, except for the segments of the critical target (i.e., an overall, monotonous, change in F_2 height and speaking rate throughout the sentence). Using an eye-tracking paradigm to study listeners' uptake of spectral and temporal cues in the conditions with manipulated vowels vs. contexts, we tested at what point in time listeners take into account the different types of context (i.e., spectral vs. temporal information). In this way we sought to gain insight into the cognitive stages at which contextual information influences the uptake of speech cues.

1.1. Contextual influences on the perception of speech cues

The influence of context sentences on speech perception has been well established (Johnson, 2005). A large number of studies has shown that temporal cues in the speech signal are interpreted relative to the speaking rate of a preceding context sentence (e.g., Kidd, 1989; Reinisch, Jesse, & McQueen, 2011a; Summerfield, 1981). To exemplify, the English sounds /b/ and /p/ are mainly distinguished by their voice onset time (VOT; i.e., the time between the release of the stop closure and the start of the vocal fold vibration in a following sound). A short VOT is associated with /b/ and a long VOT with /p/. However, a sound with an ambiguous VOT between [b] and [p] is interpreted as /p/ when it is preceded by a sentence that is spoken fast (i.e., the VOT then sounds long relative to the short segments in the fast context) but it sounds like a /b/ when the sentence is spoken slowly. As a result, context information influences whether listeners hear words as 'pin' or 'bin', to name but one example. Such influences can aid speech perception, especially in the case of ambiguous phonemes (e.g., Newman & Sawusch, 2009; Reinisch et al., 2011a; Sawusch & Newman, 2000).

A ''similar'' contextual influence has been reported for the perception of spectral speech-cues. The similarity arises out of the *contrastive* nature of the context effect in the spectral as well as the temporal domain. In the case of vowel perception, for instance, spectral properties of a preceding context have been shown to influence the location of the category boundary between phonemes such as /ɪ/ and /ɛ/ (e.g., Broadbent & Ladefoged, 1960; Sjerps, Mitterer, & McQueen, 2011a, Watkins, 1991). If a speaker's first formant (F_1) range is low, a subsequent sound which is ambiguous between [ɪ] and [ɛ] is perceived as having a relatively high F_1 and thus subjectively sounds more like /ɛ/ (which has a high F_1) than /ɪ/ (which has a low F_1). These findings, both for spectral and durational context effects align with the notion that perceptual systems operate in a contrastive way (Kluender et al., 2003).

Importantly, such contrastive influences on perception arise at a number of different stages in the perceptual system and for different modalities (e.g., Kluender et al., 2003 and references therein). For the current study we focused on what has been termed ''central'' compensation processes (Lotto, Sullivan, & Holt, 2003; Watkins, 1991). Central processes have been defined as those that take place beyond ''peripheral processing stages''. Peripheral influences have been described as those processes that take place at or before the stage of the auditory nerve, and at levels where no mayor interactions take place between information processed in the two ears (Holt, 2005; Lotto et al., 2003). It is important to prevent strong contributions from peripheral influences in the current research because these influences, either in the form of masking or as auditory enhancement (see Lotto et al., 2003, and references therein), are also contrastive in nature (Summerfield, Haggard, Foster, & Gray, 1984) and could therefore obscure the central contrastive effects under investigation here. Unlike the central compensation effects, however, peripheral influences disappear at longer intervals between two sounds or when those sounds are presented to different ears (see e.g., Holt & Lotto, 2002; Sjerps et al., 2011a; Summerfield et al., 1984; Watkins, 1991; Wilson, 1970). Therefore, in order to prevent peripheral effects in the current study, we added a short silent interval between the context sentence and the target words (see method section for more detail).

A second contrastive effect that could play a role in the experiments reported here has been described as the ''anchor-effect''. This effect relates to the observation that when one endpoint sound of a continuum (i.e., the anchor) is presented more frequently than the other sounds from the continuum, the category boundary moves towards the anchor (Sawusch & Nusbaum, 1979). Similarly, the range of acoustic values along a continuum that is presented for categorization can influence the response (Brady & Darwin, 1978). In the case of such anchor-effects the total size of boundary shifts could, in principle, be the result of contrastive effects at least two functionally different levels. The first is a perceptual stage, similar to the contrastive influences under investigation here. The second, however, is related to listeners' overt strategies to balance responses across the available response categories (note that this strategy is also contrastive). Therefore, to minimize additional influences from response strategies, the context sentences that were used here did not contain potential anchor vowels. That is, the context sentences only contained vowels other than /ɑ/ or /aː/. Any remaining strategic biases should affect our different conditions in a similar fashion and thus interference with the main question should be minimal.

The objective of the present study was to explore when during the processing of the vowel listeners interpret vowel-internal cues relative to spectral and durational context information. This should inform us at what level of processing the acoustic properties of the context can influence speech perception and help to evaluate two existing approaches concerning the locus of these context effects. The first approach suggests that both spectral and durational influences on speech perception take place at early, general auditory levels of perception (i.e., at or just preceding early cortical stages of processing). That is, the acoustic characteristics of a vowel are perceived to contrast with the context's characteristics: a high F_2 context leads to the perception of lowered F_2 target. For the spectral domain this process has been likened to an operation which is similar to applying an inverse form of a context's Long Term Average Spectrum (LTAS) to the target sound (Watkins & Makin, 1994, 1996).

Support for the early use of context information comes from findings that phoneme perception is influenced by speaking rate and spectral information from speakers other than the target speaker (speaking rate context: Green, Stevens, & Kuhl, 1994; Green, Tomiak, & Kuhl, 1997; Newman & Sawusch, 2009; Sawusch & Newman, 2000; spectral context: Lotto & Kluender, 1998; Watkins, 1991). For instance, if one speaker starts a sentence, but the sentence is finished by a different speaker, listeners' interpretation of the second speaker's speech is influenced by the auditory properties of the first speaker's utterance. These results led to claims that speaking rate and spectral information of a preceding context are used at early processing stages which precede speaker-specific adjustments (or at least early during processing acoustic context information is processed independently of speaker-specific information). Furthermore, additional support for an early implementation of context effects on speech perception comes from the observation that non-speech sounds influence the perception of subsequent speech and non-speech sounds (e.g., Holt, 2005; Sjerps et al., 2011a, and references therein). The latter suggest that some influences of context on speech sounds even precede speech-specific levels of processing.

In contrast, the second approach assumes that the influence of contextual information involves an online adjustment to a ‘‘speaker-specific frame of reference’’. This approach, termed ‘‘extrinsic vowel normalization’’ (Nearey, 1989) or ‘‘vocal tract normalization’’ (Johnson, Strand, & D′Imperio, 1999) argues that listeners estimate the formant values of a particular speaker's vowel space. This frame of reference is based on surrounding context provided by the same speaker. Subsequent speech is then adjusted to fit this specific frame of reference. This explanation attributes normalization to a ‘‘higher’’ level of processing where speaker-specific information can influence perception because the estimated frame of reference is related to a particular individual. This proposal finds support in the observation that the category boundary of a ‘‘hood’’–‘‘hud’’ (/hʊd/–/hʌd/) continuum depended on whether listeners also saw a video of a male versus a female speaker, or even merely imagined listening to a male of female speaker (Johnson et al., 1999). Since, in this case, the shift in category boundaries cannot be related to acoustic contexts (but rather to circumstantial context such as the inference about who is speaking) it is clear that higher level information can have an influence on vowel perception.

The data presented above indicate that both mechanisms (i.e., auditory-level processing and the use of speaker information) are in fact likely to play a role in speech perception to some extent. However, it is unclear whether the largest part of the perceptual shifts that have been observed due to acoustic context should be attributed to the lower-level general auditory mechanisms or to speaker-specific adjustments of the frame of reference. The current experiment can help to evaluate these hypotheses in the following way: a delay in the influence of contextual properties relative to the uptake of vowel-internal cues would provide strong evidence against a low-level auditory implementation of context effects. If, on the other hand, vowel-internal cues are immediately influenced by contextual properties (for both spectral and durational cues) this would provide suggestive evidence to the contrary, aligning with the predictions of a low-level implementation of contextual influences.

The combination of a set of recent studies provides first evidence (albeit indirect) for a low-level implementation of contextual influences on cue perception by looking at the timecourse of segment-internal vs. context effects in the domain of durational information. Shatzman & McQueen (2006) had investigated the uptake of durational cues (i.e., in a neutral speaking rate context) using an eye-tracking design. They showed that in ambiguous phrases such as Dutch *een*(*s*) (*s*)*peer* (“one/once spear/pear”) listeners are more likely to interpret the /s/ as word initial the longer its duration. Reinisch et al. (2011a) then showed that the timecourse of contextual influences on the perception of those cues (with ambiguous targets) followed a qualitatively similar pattern. That is, when the duration of /s/ was kept constant, but the ambiguous target sequence followed a fast context sentence (which made the duration of /s/ subjectively longer), listeners were more likely to interpret the /s/ as word initial than when it followed a slow context sentence. A change in subjective /s/-duration relative to the speaking rate context (Reinisch et al., 2011a) had thus qualitatively similar effects as the manipulation of actual /s/-duration (Shatzman & McQueen, 2006). To this point, however, no *direct* comparison has been made between the uptake of context information vs. sound-internal cues. Moreover, investigating the timecourse of contextual influences in both the spectral and temporal domain will further address whether the same context-to-vowel relation holds for different types of cues.

1.2. Two cues on a single phoneme: are they used at the same point in time?

The investigation of cue uptake in both spectral and durational cues across time allowed us to address an additional issue with respect to the processing of (vowel-internal) cues to phoneme identity. It is unclear to what extent the perceptual weight of speech cues such as spectral and temporal information are perceptually distributed across speech segments. That is, if two cues are present on the same speech sound, does their influence on lexical access follow the same timecourse? Previous research has shown that different cues to the same segment are used in the order they come available. For instance, VOT and duration of the following vowel are used sequentially in the perception of English stop voicing (McMurray, Clayards, Tanenhaus, & Aslin, 2008). However, in this case the vowel follows the stop, so the cues of the vowel become available later in the speech stream. As a result, the study by McMurray et al. provided a good example of how the temporal uptake of cues can be tested but it could not evaluate the timecourse of different cues that start to become available at a similar point in time. This was the second aim of the present study.

In the Dutch /ɑ/–/aː/ contrast spectral and durational information are available on the same segment (Adank et al., 2004). The current study could therefore more strictly compare potential differences in cue uptake between different types of cues—specifically spectral and durational cues. /ɑ/ is shorter and has lower first and second formants than /aː/. The question was whether, even if two cues are available in the same segment, the influence of these cues becomes evident at the same point in time. Durational cues are, by definition, spread out over time. It may thus take listeners longer to interpret durational cues than spectral cues. Spectral cues are usually already present in the form of coarticulatory cues (e.g., Dahan, Magnuson, Tanenhaus, & Hogan, 2001), and could therefore be interpretable well before the end of the vowel, possibly by the steady state part and thus earlier than durational cues. Alternatively, vowel duration could also potentially be estimated before the end of the segment. The initial part of a vowel is likely to also contain some information about vowel duration since the speed and slope of the vowel-inherent spectral change (Nearey & Assmann, 1986) are likely to be related to the overall duration of a segment. To our knowledge, no study has yet addressed the timecourse of the uptake of spectral and durational cues that are available in the same segment. The current investigation will therefore also allow for a comparison of the uptake of spectral versus durational cues to vowels over time.

Related experiments have been run with spectral cues to a vowel vs. lexical tone (Cutler & Chen, 1997). In that case, it was argued that in order to recognize the tone contour, more of the signal must be heard than is necessary for the recognition of spectral cues (but see Malins & Joanisse, 2010). For the current study we therefore hypothesized that in Dutch vowel perception, differences in duration are taken into account later than spectral information. As mentioned above, we used an eye-tracking paradigm to track the timecourse of the uptake of spectral and temporal cues in vowel perception. Since effects of different acoustic dimensions are difficult to compare we used a measure that is based on the effect-size of each cue over time, relative to its own maximum. That is, we tested at what point in time the effects of spectral and temporal information reached certain percentages (e.g., 30%, 40%, 50%, etc.) of their maxima (see e.g., McMurray et al., 2008; Miller, Patterson, & Ulrich, 1998; Ulrich & Miller, 2001; cf. methods section for details). The same analysis method was used for the main objective of the present study, the investigation of when contextual properties influence the perception of speech cues.

1.3. Experimental outline and summary of research goals

To address these questions, two experiments with two sub-parts each were constructed. For both experiments, the /ɑ/–/aː/-minimal pairs were embedded in a carrier sentence and the sentences differed either in the properties of the target vowel (Experiment 1a and 2a); in the properties of both the target vowel and the context sentences (Experiment 1b); or in only the properties of the context (Experiment 2b). The first experiment, which consisted of a simple categorization task, presented listeners with a limited range of items and had three aims. First, it confirmed the importance of both spectral and durational cues to the vowel contrast /ɑ/–/aː/ in Dutch. Second, it should reveal spectral and durational properties, which are ambiguous and therefore susceptible to context effects. Third, the experiment aimed to replicate and quantify the strength of the context effects of speaking rate and average formant frequencies on the perception of a target vowel.

Based on the results of Experiment 1, Experiment 2a used a range of F_2 frequencies and vowel durations that lead to similar changes in the perception of the target vowel as the context manipulations did. Trying to equate the effect sizes of the vowel and the context manipulations was essential since the timing of an effect can be related to its effect-size. That is, if one of two effects is much stronger than the other, it is likely that this effect becomes significant at an earlier point in time than the weaker effect. Therefore to obtain a ‘‘fair’’ comparison between the uptake of vowel-internal cues and context information we tried to match the size of the effects across Experiments 2a and 2b.

Experiment 2 examined listeners' uptake of vowel-internal vs. context cues over time using an eye-tracking paradigm. Eye tracking makes use of the fact that listeners spontaneously fixate visual referents that match the current hypotheses about the words being said (Allopenna, Magnuson, & Tanenhaus, 1998; Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; including printed words: e.g., Huettig & McQueen, 2007; McQueen & Viebahn, 2007). Importantly, eye movements are driven by the speech signal as it unfolds, making it possible to tap into the earliest moments of word recognition. Approximately 200 ms after the presentation of the speech signal, listeners direct their gaze towards potential referents that match the signal at that point in time. Note that the 200 ms lag is mainly driven by the time needed to plan and launch a saccade (Matin, Shao, & Boff, 1993) and should be stable across conditions. Using eye tracking, it was therefore possible to compare the timecourse of lexical competition between situations in which sounds can be distinguished by means of vowel-internal cues and situations in which ambiguous sounds are interpreted relative to the preceding context.

To summarize, our two main objectives were: first, to establish at what point in time during the processing of the vowel, context information is taken into account. Second, we set out to explore whether spectral and temporal cues are taken into account simultaneously or whether temporal cues are processed more slowly since they require more time to unfold before they can be properly evaluated.

2. Experiment 1

Experiment 1 was set up to replicate previous findings that Dutch listeners use spectral and durational cues to distinguish the Dutch /ɑ/–/aː/ vowel contrast. Moreover it should demonstrate that both spectral and durational cues to the vowel are influenced by the spectral and durational characteristics of a context sentence.

Experiment 1 was split up into two parts, Experiment 1a and Experiment 1b. In Experiment 1a, participants categorized minimal word pairs with vowels manipulated along continua varying on two dimensions: a five-step second formant (F_2) frequency continuum crossed with a five-step duration continuum (i.e., resulting in 25 combinations). This allowed us to test the use of both cues and to select perceptually ambiguous vowels.

Experiment 1b tested the susceptibility of spectral and durational cues in the vowel to context effects. Participants again categorized minimal word pairs differing in the /ɑ/–/aː/ contrast. However, now the words contained vowels from only the ambiguous part of the continuum (3 steps along the spectral and durational dimension). Importantly, now the words were presented in spectrally and durationally manipulated context sentences. The context manipulations consisted of two spectral values of F_2 (high vs. low), crossed with two duration values or speaking rate values: fast–slow. (Note that speaking rate was manipulated in a linear fashion and thus the term is used interchangeably with duration; duration will be used with reference to the vowel manipulation, rate with reference to the context manipulation).

This experiment provided two additional pieces of information: first, because the vowel stimuli used in Experiment 1b consisted of a subset of those in Experiment 1a it provided a more fine-grained picture about which vowel tokens could be considered ambiguous. This most ambiguous vowel was used in Experiment 2b. Second, Experiment 1b provided an estimate of the size of the perceptual shift caused by spectral and durational context. These estimates were then used to manipulate the vowel-internal cues for Experiment 2a so that they matched the size of the context effects in Experiment 1b (and later the expected effects in Experiment 2b).

2.1. Methods

2.1.1. Participants

Sixteen participants from the Max Planck Institute for Psycholinguistics participant pool were tested; eight participated in Experiment 1a and eight in Experiment 1b. They were native Dutch speakers and received a monetary reward for their participation. None of them reported hearing disorders.

2.1.2. Materials: Experiment 1a

Twenty-eight Dutch monosyllabic minimal word pairs were selected for the complete series of experiments. One word of a pair contained the vowel /ɑ/ and the other word contained the vowel /aː/, for example, *tak–taak* (/tɑk/–/taːk/; ‘‘branch’’–‘‘task’’; see Appendix A for a list of all word pairs). The chosen word pairs had a variety of simple and complex syllable-onsets and codas but consonants were restricted to obstruents and nasals. Approximants, and especially liquids, tend to color the vowel to an extent that would have affected the splicing procedure (West, 1999; for details about the splicing procedure see below).

A male Dutch native speaker recorded both words of the twenty-eight minimal pairs in a sound-conditioned booth, three times each. Words were read embedded in the carrier sentence *Klik nu een keer op het woord* [*TARGET*] *boven de ster/driehoek/rechthoek/cirkel* (‘‘Now click once on the word [TARGET] above the star/triangle/rectangle/circle’’). Note that the context following the target was not informative for completing the task in any of the experiments but was included to maximize the amount of context information listeners could possibly use. Each target was drawn from the set of minimal pairs. Measurements of the vocalic portions of the target words (excluding transitions from the consonants) were taken to estimate the distribution of spectral and durational properties of the two vowels as spoken by the speaker. Measurements included the first two formants (F_1 and F_2: mean value of the steady-state portion of the vowel) and duration. Based on these measures (and their subjective naturalness after manipulation, judged by the second author who is a native speaker of Dutch), three vowel tokens (out of the 84 measured) were

Table 1
Durations (in ms) and F_1 and F_2 frequencies (in Hz, measured from the steady-state portion of the vowel) for the /aː/ vowels that were selected for manipulation (indicated by *). Values of the /ɑ/ counterparts of the minimal pairs are given for comparison, as are values for the preceding and following contexts. Formant estimates for words and context sentences were based on vocalic portions only.

Source	Transition-type	Duration	F_1	F_2
aas*	glottal/velar	191	653	1349
smaak*	labial	149	645	1343
staat*	alveolar	154	622	1344
as	glottal/velar	103	603	1068
smak	labial	83	627	1087
stad	alveolar	106	533	1089
Preceding context	–	1220	372	1400
Following context	–	825	379	1312

Note that in some cases the F_1 values for the minimal pairs are relatively close (as in the labial pair), which seems surprising given the typical F_1 difference for the Dutch /aː/–/ɑ/ contrast. However, these comparisons are based on single instances. Moreover, the /aː/ tokens were partly selected based on how natural their resynthesis to /ɑ/ sounded. It is likely that this is why tokens with a relatively low F_1 were selected as the result would lead to a naturally sounding /ɑ/.

selected for further manipulation. Three, rather than one vowel token, were selected due to the different places of articulation of the preceding consonants (i.e., glottal/velar, labial, and alveolar; note the subdivision in Appendix A). The three selected vowels were taken from recorded instances of *smaak* (/smaːk/, ''taste'', labial transition), *staat* (/staːt/, ''state'', alveolar transition) and *aas* (/ʔaːs/, ''bait'', transition from a glottal stop which was used for velars). This way, appropriate sounds could all be spliced into minimal pairs without introducing noticeable acoustic artifacts (i.e., vowels were matched with the transition of the onset consonants; note that for the velars the particular vowel with transitions from a glottal stop sounded most natural, hence we decided to use this token for the velar contexts). The three vowel tokens will be referred to as transition-types. Transition-type was included as a factor in all statistical analyses but will only be reported for cases where it reached significance. The selected transition-type vowels fell within 0.8 standard deviations of the mean on all acoustic measures across all recorded instances of /aː/ and were manipulated separately. The top rows of Table 1 display the duration and F_1 and F_2 values for the three /aː/-vowels selected for manipulation and values of a selected instance of the /ɑ/ counterparts for purposes of comparison. Note that long vowels were selected as a starting point for manipulation, since cutting out pitch periods for the duration manipulation introduces fewer artifacts than duplicating pitch periods.

The three selected transition-type vowels were manipulated for duration and spectral cues using PRAAT software (Boersma & Weenink, 2009). We only manipulated F_2, since in informal pretesting F_2 proved to be sufficient for the distinction of the /aː/–/ɑ/ contrast. F_1 was kept at a fixed value, across all steps of the continuum. We chose to only manipulate one formant because, if spectral context effects are operating at a general auditory level of processing, then the influence of context on a target F_2 might not be restricted to the frequency range of F_2 in the context sentence but also the values of F_1 (and vice versa). As such, the influence of F_1 and F_2 could potentially interfere with each other.[1] Therefore we decided to manipulate only F_2, both in the vowels and in the context. Vowel continua from [ɑ] to [aː] were created by manipulating the factors Spectrum (F_2) and Duration (later used interchangeably with speaking rate when it comes to context manipulations in Experiment 1b).

The factor Spectrum was manipulated along a 5-step continuum of F_2 values from [aː] to [ɑ]. The F_2 manipulation was based on Burg's LPC method, in analogy to the manipulation reported in Sjerps et al. (2011a). The source and filter models were estimated automatically from the selected instances of /aː/. The F_2 values in the filter models were inspected and increased or decreased to form a continuum. The continuum ranged from 100 Hz above the base vowels' original F_2 values to 200 Hz below the original values, that is, the continuum spanned approximately from 1145 Hz to 1445 Hz (with slight variation between transition-types as indicated in Table 1). Thus the step size was 75 Hz. Then, the source and filter models were recombined and the new sounds were adjusted to have the same amplitude envelope and the same overall amplitude as the original recordings. The adjustment for the factor duration consisted of creating a 5-step duration continuum on each of the spectrally manipulated vowels. Target vowels were shortened or lengthened manually by removing or duplicating individual pitch periods throughout the vowels. The duration continuum spanned 80 ms, from 100 ms to 180 ms, in steps of approximately 20 ms as permitted by the duration of the individually removed periods. The duration values on the continuum fell within the range of /aː/ and /ɑ/ tokens as naturally produced by our speaker (see Table 1).

For Experiments 1a and 1b, only three words of the full set of recorded minimal pairs were used. The consonantal frames were different words than those from which the vowels were originally taken. This was to anticipate the situation of Experiment 2 where the three selected vowels were spliced into the whole set of consonantal frames of the minimal pairs. The vowel with glottal/velar transitions was spliced into the consonant frame of *gaas* (/ːχaːs/, ''gauze''; henceforth only referred to as ''velar'' transition), the vowel with alveolar transitions was spliced into *zak* (/zɑk/, ''bag''), and the vowel with labial transitions was spliced into *maagd* (/maːχt/, ''virgin''). The words were then each spliced into one selected carrier sentence. The carrier sentence was selected for having a duration that was close to the average duration of all recorded carrier sentences and for having an intonation contour that sounded natural (i.e., no pitch jumps) when the target words were spliced in. For Experiment 1a, this carrier sentence was used without further manipulation. A 300 ms gap was introduced between the context and the target words to prevent peripheral contrastive effects (see the section ''Introduction''). The precursors were spoken with a prosodic break before the target (which was produced as a phrase of its own) such that this silent interval did not sound particularly unnatural.

2.1.3. *Procedure: Experiment 1a*

Each combination of the 5 spectral (F_2) steps × 5 duration steps × 3 transition-type vowels (i.e., the three selected words) was presented four times to each participant, for a total of 300 trials. All stimuli were presented in random order such that all stimuli were presented once before a repetition occurred. On each trial, participants saw two words of a minimal pair appearing on the screen with the word containing /aː/ always on

[1] To exemplify: a high F_2 context should lead to more /ɑ/ responses as it pushes the perception of the ambiguous sound's F_2 ''downwards''. However, the rising F_1 could potentially have the opposite effect on the perception of the target F_2 as it enters the frequency region of F_2.

their right (e.g., 'Gas' 'Gaas'). Although the constant target position could trigger a preference or bias for one of the responses, the "handicap" applied equally to all conditions and can thus not explain the effects of the individual cues. 500 ms after the onset of the display, listeners heard the stimulus sentence, presented at a comfortable listening level. The listeners' task was to select the word they perceived. Responses were given by pressing the left or right button on a computer mouse using either their two thumbs or index fingers. Participants were tested in a sound-conditioned booth, wearing Sennheiser HD 280-13 headphones. The experiment was run using Presentation software (Version 11.3, Neurobehavioural Systems Inc.). Every 50 trials, participants were allowed to take a self-paced break.

2.1.4. Results: Experiment 1a

Analyses of the responses were carried out using linear mixed-effects models (Baayen, Davidson, & Bates, 2008) as provided in the lme4 package (Bates & Sarkar, 2007) in R (version 2.10.0; The R foundation for statistical computing). Particularly for the binomial data at hand this analysis method has been argued to be superior to traditional ANOVAs, as it capitalizes less on Type-I errors (Quené & Van Den Bergh, 2008). For the dichotomous dependent variable of click responses (/ɑ/=1 vs. /aː/=0), the logit linking function was used (Jaeger, 2008). Data will be reported with beta scores, indicating the direction of the effects. A *z*-score is provided for the proportion correct data (based on Wald's *z*-score) indicating the distance of the coefficient from zero in terms of its standard error (Jaeger, 2008). For Experiment 1a, responses were analyzed by fitting models with Spectrum, Duration/Rate, and Transition-type (along with their interactions) as fixed factors, and participant as a random factor. Furthermore all within-participant factors were entered as random slopes into the random effects structure (if interactions between these factors were also added to the random effects structure the models failed to converge). Note that "item" was not included as a random factor in Experiment 1 since only three different items were used and thus a "by-item" analysis would not be meaningful. The best model was established through a backward elimination procedure, such that non-significant predictors were removed from the model. The models were then compared by means of a likelihood ratio test.

The numerical fixed factors Spectrum and Duration/Rate were coded such that they ranged from −0.5 (low F_2 and fast/short) to 0.5 (high F_2 and slow/long) at their endpoint values. Negative regression weights therefore indicated more /ɑ/-responses corresponding to shorter vowels and vowels with lower F_2. For the factor Transition type (levels: velar (*gaas*), labial (*maagd*), alveolar (*zak*)) the level *velar* was modeled as the reference level (i.e., "gaas" comes first in the alphabetical order of the three words used in the experiment). That is, all other effects (i.e., Spectrum, Duration/Rate, and their interaction) were first tested on the level of Transition type that formed the reference variable (i.e., velar). A potential interaction of the factor Transition type with other factors then indicates a difference of these factors' effects across the transition-types (e.g., an interaction of Transition type at the alveolar level with the factor Spectrum indicates that the effect of Spectrum differs between the velar level of the factor Transition-type and the alveolar level). A lack of an interaction between the factor Transition-type and the other factors therefore shows that the respective effect does not differ across the transition-types and can as such be interpreted to apply equally to all stimuli.

The left panel of Fig. 1 displays the categorization results as mean proportions of /ɑ/-responses for each of the 25 stimuli from the vowel grid. The amount of black fill for each of the boxes provides a visual display of these proportions. The right panel displays mean reaction times (RTs) to each of the stimuli. Measures are given from the onset of the target words but the bars display reaction times only within the range from 600 to 1100 ms (to visually enhance the pattern; all averaged values were between 600 and 1100 ms). RTs are plotted for illustration only. They show the expected pattern that responses to ambiguous sounds take longer on average than responses to unambiguous sounds. Statistical analyses were carried out for proportions of /ɑ/-responses. The analyses showed that, for the velar transition-type words, listeners had a preference for /aː/-responses ($b_{intercept}=-1.474$, $z=-3.843$, $p<.001$). The effect of Transition type, however, indicates that the overall preference for /aː/-responses was less strong for the labial transition type and absent for the alveolar transition type (compared to the velar transition-type; $b_{labial}=0.669$, $z=2.536$, $p=0.011$; $b_{alveolar}=1.596$, $z=3.722$, $p<.001$). The effect of Spectrum indicates that velar tokens with a low F_2 were perceived as /ɑ/ more often than tokens with a high F_2 ($b_{spectrum}=-5.751$, $z=-8.511$, $p<.001$), and the effect of Duration/Rate shows that short tokens were perceived as /ɑ/ more often than long tokens ($b_{rate}=-5.096$, $z=-8.165$, $p<.001$). The interaction between Spectrum and Duration/Rate reflects the fact that the effect of F_2 was slightly larger for the longer the items ($b_{spectrum \times rate}=-2.358$, $z=-3.291$, $p<.001$). Two more interactions were found: An interaction was found between Transition type and Duration/Rate ($b_{rate \times alveolar}=-3.0522$, $z=-4.637$, $p<.001$; $b_{rate \times labial}=-2.0322$, $z=-3.338$, $p<.001$). This reflects the fact that the effect of Duration/Rate was even stronger for the alveolar and labial

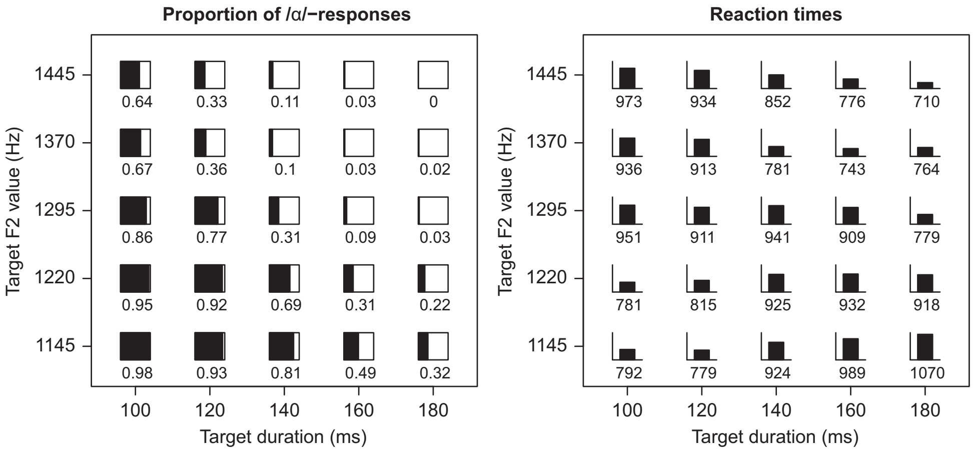

Fig. 1. Experiment 1a. The left panel displays proportions of /ɑ/-responses to the duration- and F_2-based vowel continua, collapsing across transition types. The *x*-axis displays mean vowel durations for the different duration steps (in milliseconds). The *y*-axis displays mean F_2 values for the different steps along the spectral continuum (in Hz). For each combination of duration and F_2 value, the amount of black fill of the box reflects the proportion of /ɑ/-responses (the value is indicated below each box). The right panel displays reaction times and is organized the same way as the left panel. For each combination of duration and F_2, the length of the bar indicates the mean RT from target word onset (also printed below the graph). RTs are plotted in the range from 600 to 1100 ms to enhance visibility of the pattern. All values are overall mean values and may thus slightly differ for the three individual transition types as given in Table 1.

transition types than for the velar transition type. The other interaction involved Transition type and Spectrum ($b_{spectrum \times labial}=-0.2318$, $z=-0.388$, $p=0.698$; $b_{spectrum \times alveolar}=1.735$, $z=2.993$, $p=0.003$). It indicated that the effect of Spectrum was slightly smaller for the subset of alveolar transition types when compared to the velar ones. This shows that in some cases a similar acoustic change can have perceptual effects of different sizes. However, the three transition types consisted of different tokens. What Experiment 1a should ensure is that all selected and manipulated vowel tokens are categorized based on spectral as well as durational cues. This can be seen in Fig. 1. Note that the effect of Rate seemed slightly larger per step size on our continuum.

2.1.5. Materials: Experiment 1b

Based on the results of Experiment 1a, ambiguous values of the vowels were selected for the three transition-type tokens. These values were (in the order Spectrum, Duration) 1224 Hz, 140 ms for the velar transition-type vowel, 1294 Hz, 130 ms for the alveolar transition-type vowel, and 1256 Hz, 130 ms for the labial transition-type vowel. These values were now taken to define the midpoint of a fine-grained grid that consisted of three steps for each of the two factors Spectrum and Duration/Rate (3 duration steps crossed with 3 spectral steps). For the factor Spectrum, the grid spanned F_2 values of 75 Hz around the selected ambiguous values (37.5 Hz in both directions from the selected ambiguous sound). For the factor Duration/Rate, the grid spanned 20 ms around the selected values (10 ms in each direction, as permitted by individual pitch periods). Note that the step sizes here are half of the step sizes in Experiment 1a. The vowel manipulation followed the same procedures as in Experiment 1a. The vowels were then spliced in the same three words.

In addition, four versions of the carrier sentence were created. The spectral manipulation of the contextual speech materials was carried out in a similar fashion as was done for the vowels using the LPC method. The manipulation involved an increase or decrease of F_2 in all vowels by 200 Hz (after Watkins & Makin, 1994). The spectrally manipulated sentences were then linearly compressed or expanded using PSOLA as implemented in PRAAT (Boersma & Weenink, 2009). The fast version of the sentence was set to 66% of the original duration, and the slow version was set to 133% of the original duration. These values were based on previous studies in which these values showed robust rate effects on the perception of Dutch ambiguous word boundaries and durational cues to lexical stress (Reinisch et al., 2011a, 2011b). The top panel of Fig. 2 displays the LTAS (bin width = 10 Hz) of the manipulated endpoint vowels (the long tokens only because the LTAS should differ minimally across the duration manipulation). The bottom panel of Fig. 2 displays the LTAS of the manipulated context sentences for the spectral contrast. The *x*-axis is logarithmic. It can be observed that /ɑ/, the low F_2 target, has more energy than /aː/, the high F_2 target, at low frequencies. Similarly, it can be observed that the Low F_2 precursor has more energy than the high F_2 precursor at low frequencies. Importantly, the differences between the LTAS of the precursors are in roughly the same frequency regions as the differences between the LTAS of the endpoint targets. This means that those frequencies that listeners could use to distinguish between the /ɑ/ and /aː/ are roughly the same frequencies that constitute the acoustic difference between the precursors in the High and Low F_2 conditions. This means that a purely auditory contrastive compensation process should result in reliable shifts in perception. Carrier sentences were again combined with target words, leaving a silent gap of 300 ms before and after the target, for the same reasons outlined above.

2.1.6. Procedure: Experiment 1b

4 repetitions of each of the 3 vowel-spectrum steps × 3 vowel-rate steps × 2 context-spectrum steps × 2 context-rate steps × 3 transition-type vowels were presented for a total of 432 trials. All stimuli were presented in random order such that all stimuli were presented once before a repetition occurred. The experimental setup was the same as that in Experiment 1a.

2.1.7. Results: Experiment 1b

Statistical analysis procedures were identical to that in Experiment 1a, with the addition of the two context factors. Note, however, that in Experiment 1b, the interpretation for the regression weights of the vowel continua is identical to that of Experiment 1a, but the opposite holds for context effects. Context effects are contrastive; that is, a faster context with a lower F_2 should lead to fewer /ɑ/-responses. Fig. 3 displays a breakup of the proportions of /ɑ/-responses for each step of the continua in the four different context conditions. Each panel displays the values for the different vowel steps within a given context condition. The results show that the F_2 and the duration of the vowels as well as of the context

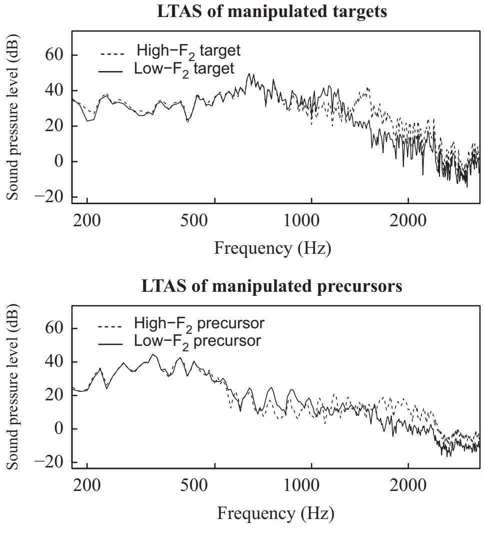

Fig. 2. Top panel: Long Term Average Spectra (LTAS) of the endpoint targets in Experiment 1a; Bottom panel: LTAS of the manipulated context sentences in Experiment 1b (see text for details).

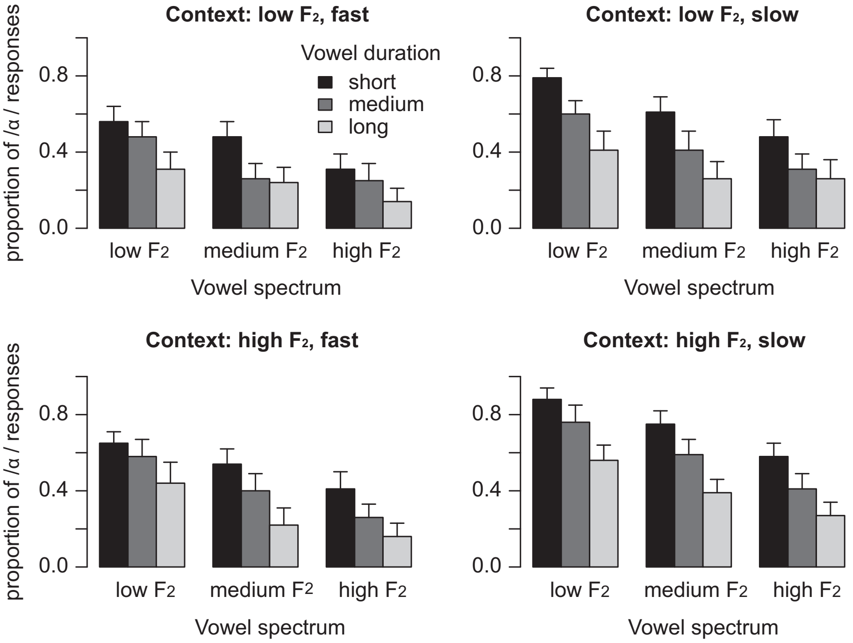

Fig. 3. Each panel displays the values for the different vowel steps within a context condition. The different panels display different context conditions. Panels on the left display categorization in the fast context condition, panels on the right in the slow context condition. Top panels display data obtained in the low F_2 condition, bottom panels display data from the high F_2 condition. Error bars reflect standard errors for convenience of reading the graphs (but note that the statistical analysis was based on linear mixed-effects models).

contributed to vowel identification. The shorter the vowel and the lower its F_2, the more /ɑ/-responses were given. For the contexts, the condition with low F_2 and faster rate led to fewer /ɑ/-responses than the condition in which the context had a high F_2 and was spoken slowly. The context combinations which induce conflicting percepts (low F_2, slow and high F_2, fast) show intermediate proportions of /ɑ/-responses.

These observations were confirmed by statistical analyses. The model included five fixed factors and their interactions as well as participant as random factor with random slopes for all factors. To ensure convergence of the models, random slopes were again fitted without interactions. Fixed factors were: Transition-type with 3 levels: velar (''gaas'' was mapped onto the intercept as reference level), labial (adjustment relative to the reference level), alveolar (adjustment relative to reference level); Vowel-spectrum (3 levels: low $F_2=-0.5$, mid $F_2=0$, high $F_2=0.5$), Vowel-rate/duration (3 levels: short$=-0.5$, mid length$=0$, long$=0.5$), Context-spectrum (2 levels: low $F_2=-0.5$, high $F_2=0.5$) and Context-rate (2 levels: fast$=-0.5$, slow$=0.5$). The analysis revealed main effects for the vowel as well as the context manipulations. The effect of Vowel-spectrum indicates that at the reference level sounds with a low F_2 were more often interpreted as /ɑ/ than words with a high F_2 ($b_{vowel\text{-}spectrum}=-1.302$, $z=-5.075$, $p<.001$). The effect of Vowel-duration/rate indicates that short sounds were more often interpreted as /ɑ/ than long vowels ($b_{vowel\text{-}rate}=-1.882$, $z=-8.447$, $p<.001$). Context effects indicated that more sounds were interpreted as /ɑ/ in a context with a high F_2 (49.0%) than with a low F_2 (39.8%; $b_{context\text{-}spectrum}=0.638$, $z=4.579$, $p<.001$), and more sounds were interpreted as /ɑ/ in a slow context (51.7%) than in a fast context (37.1%; $b_{context\text{-}rate}=0.985$, $z=5.564$, $p<.001$).

These effects were complemented by an effect of the factor Transition type ($b_{transition\text{-}type(alveolar)}=1.737$, $z=2.602$, $p=0.009$), indicating that the vowels with preceding alveolar sounds led to more /ɑ/-responses than when preceded by velar sounds. An interaction was found between Vowel-spectrum and Transition type ($b_{vowel\text{-}spectrum \times transition\text{-}type(labial)}=-1.153$, $z=-4.286$, $p<.001$) indicating that the effect of Vowel-spectrum was stronger for the vowel in the word ''maagd'' with labial transitions than with the velar words. Each of the three transition types, however, induced effects of Vowel-spectrum. The lack of an interaction between the factor Transition-type and the other factors of interest indicates that the reported effects apply to each of the three vowels to a similar extent. Finally, the optimal model also included a three-way interaction between Context-rate, Target-duration and Target-spectrum ($b_{context\text{-}rate \times vowel\text{-}duration \times vowel\text{-}spectrum}=1.161$, $z=2.082$, $p=0.037$). We do not have a clear interpretation of this effect, and given the relatively small effect size we will not further discuss this three-way interaction.

2.2. Discussion

Experiment 1a showed that both the duration and the F_2 values of the vowels strongly influenced participants' categorizations of the vowels along the /ɑ/–/aː/ continua. This is in line with previous findings (Escudero et al., 2009; Gerrits, 2001; Nooteboom & Coden, 1984; van Heuven et al., 1986) and was replicated in Experiment 1b. Importantly Experiment 1b showed that listeners' categorization responses were also strongly influenced by the speaking rate and F_2 level of the carrier sentences. Experiment 1 thus established that Dutch listeners use spectral and temporal cues in the vowel as well as in the context to identify vowels along an /ɑ/–/aː/ continuum. Experiment 1 further provided the basis for the time-course investigation of the uptake of vowel-internal cues and context information in Experiment 2.

3. Experiment 2

Experiment 2 was set up to test the timecourse of Dutch listeners' use of spectral and durational/speaking-rate cues, when they are available within the vowel vs. when the uptake of vowel internal cues is influenced by contextual information. Experiment 2 paralleled Experiment 1 in that in Experiment 2a we addressed the uptake of vowel-internal cues and Experiment 2b the context effects were central. The results of Experiment 1 were used to select the ambiguous vowels for Experiment 2b. Additionally they served to address the concern that in order to be able to compare effects of different ‘‘cue-locations’’ (i.e., vowel-internal cues vs. context information) over time it was necessary to roughly equate the expected effect sizes of vowel-internal cues and the effect of the contextual influences. Therefore, we established the relation between the strength of the context effect (defined in terms of the shift in proportion /ɑ/ responses) when compared to the strength of effects of the vowel-internal cues. To this end, the differences in vowel-internal cues between /ɑ/ and /aː/ in Experiment 2a were modeled on the size of the perceptual shifts triggered by the contexts in Experiment 1b. That is, results from Experiment 1b were used to estimate how big the acoustic shift in the targets of Experiment 2a should be in order to mimic the size of the shift in vowel perception resulting from the context information. More detail on this procedure is provided in Appendix B.

3.1. General method

3.1.1. Participants

Sixty-eight participants who had not taken part in Experiment 1 participated for monetary compensation. Half completed Experiment 2a, and the other half completed Experiment 2b. All participants reported normal hearing and had normal or corrected-to-normal vision.

3.1.2. Materials

Based on the data from Experiment 1, the most ambiguous parameters (F_2 and duration) for vowel tokens per transition-type were identified. The choice was based on the vowels’ ambiguity, and specifically on its susceptibility to context effects in Experiment 1b. The most ambiguous values were 1186 Hz, 139 ms for the vowels with velar transitions, 1294 Hz, 129 ms for the vowels with alveolar transitions, and 1256 Hz, 122 ms for the vowels with labial transitions. For Experiment 2a, these values were used to create four new targets around each vowel token: a short vowel with a high F_2, a long vowel with a high F_2, a short vowel with a low F_2, and a long vowel with a low F_2. The values were chosen to match the effect sizes to be expected in Experiment 2b, based on the results of Experiment 1b. See Appendix B for a description of the calculation of these parameters.

The four manipulated vowel versions were then spliced into the appropriate preceding and following consonants to recreate the full set of recorded target words (28 in total, see Appendix A). Transition types were matched according to the place of articulation of the consonant preceding the vowel. Each target word thus appeared in four vowel conditions. The words were then spliced into the neutral sentence context that had been used in Experiment 1a (see the bottom two lines of Table 1 for duration and mean formant frequency measures of the carrier sentence). For Experiment 2b, the selected ambiguous vowel tokens were spliced into the consonant frames of the complete set of twenty-eight minimal pairs without further manipulation. Note that this is in contrast to Experiment 1b where the vowel was also manipulated in F_2 and duration. Again, the vowels’ formant transitions matched the place of articulation of the consonant preceding the vowel. The twenty-eight target words were then inserted into the four manipulated sentence contexts used in Experiment 2b. To exclude possible differences between Experiments 2a and 2b due to the presence vs. absence of manipulation artifacts in the context sentences, the neutral context in Experiment 2a was also subjected to spectral and temporal manipulations. The manipulation, however, was implemented such that the ultimate amount of change was zero. That is, the speaking rate was speeded up and then slowed down again to match its original duration. The same was done with F_2 frequency. The manipulation thus only introduced possible manipulation artifacts. As in Experiment 1, silent periods of 300 ms were inserted preceding and following the target word.

3.1.3. Procedure

The procedure was identical for Experiment 2a and 2b. In the course of the experiment, each word pair was presented to each participant with all four cue combinations (vowel-internal cues in Experiment 2a and manipulated context sentences in Experiment 2b). The order of words was randomized with the restriction that all 28 word pairs were presented once before a repetition occurred. Within each of these four repetition blocks one quarter of the word pairs appeared with each of the four cue combinations. Across participants each word occurred equally often in each cue combination and in each block.

Participants were tested individually in a sound-conditioned booth. They were seated approximately 60 cm in front of a 32.5 cm by 25 cm screen and fitted with a head-mounted Eyelink II eye-tracking system (SR Research). Eye movements were recorded using pupil-tracking at a rate of 250 Hz. The minimal word pairs were presented orthographically in lower-case Lucida Console font, size 30, centered in the left and right halves of the screen (see Fig. 4). Printed words rather than pictures of the referents were used to be able to use a large set of minimal pairs contrasting in the /ɑ/ vs. /aː/ vowels. Picture referents would have imposed severe limitations on the use of words due to problems in depictability. Importantly printed words have been shown to function well as referents in tracking lexical competition on the phonological level of processing (see e.g., Huettig & McQueen, 2007; McQueen & Viebahn, 2007). The side of /ɑ/ vs. /aː/ words on the screen was counterbalanced across words, conditions, and participants. The average four-letter word spanned 3.2 cm, or a visual angle of approximately 3.06 degrees. The words were located above one of the geometrical shapes mentioned in the recordings (i.e., circle, star, triangle, rectangle). The shape always matched the audio. Importantly, on every trial both words of a pair were combined with the same shape making it uninformative for the completion of the task.

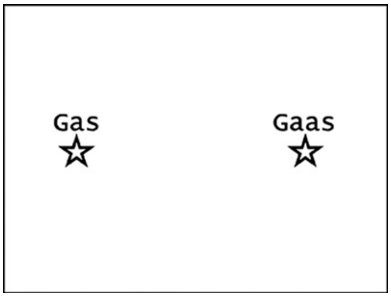

Fig. 4. Example display of the visual stimuli presented in Experiment 2.

The shapes had side lengths of size approximately 1–1.5 cm (see Fig. 4) and were included in the visual display to provide complete reference for the auditory instructions. Participants' task was to listen to the audio and click with the computer mouse on the word they thought they heard. Recording click responses in addition to the eye movements allowed us to assess the final stage of the recognition process and compare it to the implicit updating of lexical hypotheses over time as measured by eye tracking.

On each trial, listeners first saw a fixation cross for 1 s in the middle of the screen to focus listeners' attention and center their eye gaze. The cross was immediately followed by the two words and shapes. 500 ms later, the auditory stimuli were presented. This brief delay allowed listeners to scan the screen and identify on which side the word with /ɑ/ and /aː/-vowel were presented. The words and shapes stayed on the screen until the participant responded by clicking with the computer mouse on one of the words. Following an inter-trial-interval of 1 s, the next trial started automatically. After every 10th trial, a drift correction was carried out in order to adjust for possible shifts of the head-mounted eye tracker and thus a shift of the previously calibrated reference points. The experiment was controlled by Experiment Builder software (SR Research). It consisted of 112 trials and took approximately fifteen minutes to complete.

3.2. Results

Two participants from each of Experiment 2a and 2b used only one response option all the time, suggesting that they were entirely insensitive to our manipulations. On this basis their data was excluded from all further analyses. Only trials in which participants clicked in the vicinity of the word or shape (radius of 140 pixels which approximates to 4.4 cm) were analyzed. Only three trials in Experiment 2a and one trial in Experiment 2b had to be excluded based on this requirement. Analyses were run for listeners' click responses as well their eye-movement behavior. Note again that since context effects operate contrastively (e.g., a fast context makes the following vowel sound longer) the effects of vowel-internal cues and the effects of the context are expected to be in opposite directions.

3.2.1. Click responses

Statistical analyses of click responses were identical to the analyses of Experiment 1. Listeners responded on average 1704 ms after vowel onset in Experiment 2a and 1725 ms after vowel onset in Experiment 2b. Despite the small vowel manipulations in Experiment 2a, participants used spectral cues and duration (i.e., factors Vowel-rate and Vowel-spectrum) to determine their click responses. The left part of Table 2 shows percentages of participants' click responses in each of the four vowel conditions. A short vowel led to more /ɑ/-responses than a long vowel. Similarly, a vowel with a low F_2 was perceived as /ɑ/ more often than a vowel with a high F_2 ($b_{vowel\text{-}spectrum}=-0.697$, $p<.001$; $b_{vowel\text{-}rate}=-0.990$, $p<.001$; $b_{vowel\text{-}spectrum \times vowel\text{-}rate}=0.082$, $p=0.63$).

In Experiment 2b, listeners gave the most /ɑ/-responses when the context was slow and had a high F_2 (see right side of Table 2). Following a slow rate context, the ambiguous vowel sounded relatively shorter than when it followed a fast context. This led to more /ɑ/-responses. Similarly, following a high F_2 context, the second formant of the vowel sounded relatively lower than when it followed a low F_2 context. This led to more /ɑ/-responses in the high F_2 context than in the low F_2 context. Statistical analyses confirmed these observations ($b_{context\text{-}spectrum}=0.921$, $p<.001$; $b_{context\text{-}rate}=1.359$, $p<.001$). Additionally, an interaction was observed between Spectrum and Rate ($b_{context\text{-}spectrum \times context\text{-}rate}=0.413$, $p<.05$). The interaction suggests that the effect of Spectrum is larger for the slow context sentences than the fast context sentences. Note that the slow sentences are per definition longer than the fast sentences hence the spectral effect may be larger due to the increased length of the context (see e.g., Holt, 2006).

These findings replicate Experiment 1 and show that our manipulations were successful as participants used spectrum as well as rate to distinguish the /ɑ/–/aː/ contrast. Cues could be interpreted when present vowel-internally (Experiment 2a) as well as from the context when cues on the vowels were ambiguous (Experiment 2b).

3.2.2. Eye-tracking

3.2.2.1. Overall analyses. If the gaze on a word fell within a predefined circle of radius 4.4 cm around the middle of the word-shape group, it was counted as a fixation. Participants' fixations on the /ɑ/-word in Experiments 2a and 2b are shown in Fig. 5. The figure shows fixation proportions on the /ɑ/-words depending on the spectral and temporal values in the vowel (i.e., long vs. short duration and low vs. high F_2; left panel) and in the context (slow vs. fast rate and low vs. high F_2, right panel). Fixation proportions are plotted from the acoustic onset of the vowel. The outer solid vertical lines show the time window of analyses from 200 ms to 820 ms after vowel onset. This time window spans the time from vowel onset until the average onset of the following context shifted by 200 ms. A time lag of 200 ms is commonly used as an estimate of the time needed to program and launch a saccade (see, e.g., Allopenna et al., 1998). That is, on average, eye movements start reflecting word recognition 200 ms after the onset of the target word. The time window thus spanned from the earliest moment during which processing of the vowel was reflected in the eye movements until the point in time at which further context information could add to the effect.

For the overall analysis, fixation preferences on /ɑ/-words over /aː/-words were calculated. First fixation proportions on the /ɑ/-words and /aː/-words were logistically transformed and then subtracted (see Barr, 2008; Jaeger, 2008, for discussions of the analysis of untransformed data). Linear mixed-effects models were run with participants' preference to fixate on the /ɑ/-word over the /aː/-word as the dependent variable

Table 2
Percent click responses on the /ɑ/-word in each of the four conditions in Experiments 2a and 2b. Note that in Experiment 2a the manipulations were applied to the target vowels whereas they were applied to the contexts in Experiment 2b.

Spectrum	Rate			
	Experiment 2a (vowel-internal manipulation) (%)		Experiment 2b (context manipulation) (%)	
	Short	Long	Fast	Slow
Low F_2	66.8	50.6	44.1	60.9
High F_2	55.2	39.6	54.9	74.4

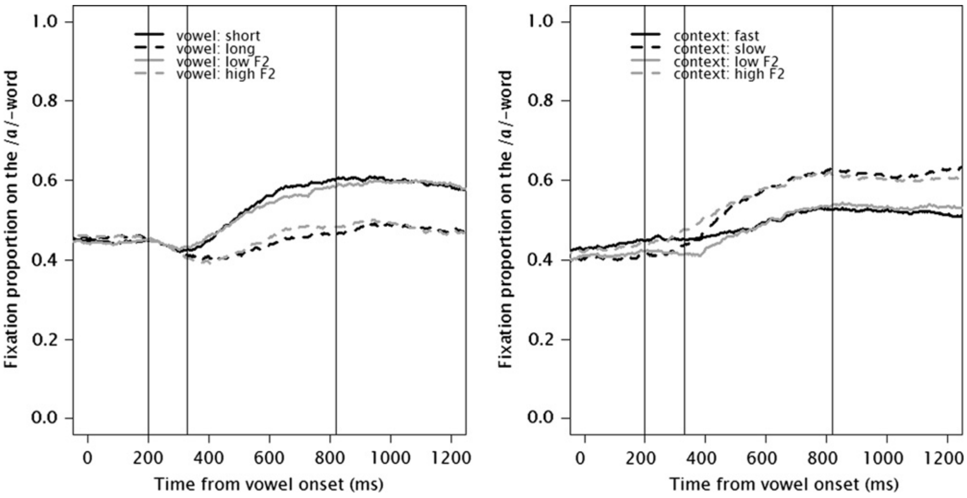

Fig. 5. Fixation proportion on the /ɑ/-word over time from acoustic vowel onset in Experiments 2a (vowel-internal cues) and 2b (context manipulation). The outer vertical lines show the limits of the overall time window (see text for details), the ‘‘middle’’ vertical line indicates vowel offset shifted by 200 ms. The left panel shows fixation proportions for each cue in Experiment 2a, and the right panel shows fixation proportions in Experiment 2b. The plots show the contributions of each of the two levels of the spectral (i.e., high vs. low F_2) and temporal (long vs. short/fast vs. slow) manipulation.

and Spectrum, Rate, and the interaction of the two as fixed factors, as has been done in Experiment 1. For the random effects structure, random intercepts were fitted for participant and word pair with additional random slopes for the within-participant and within-word pair factors Spectrum, Rate, and their interaction. Spectrum and Rate were effect-coded such that effects of Spectrum and Rate can be interpreted as main effects. Due to the inclusion of random slopes the output of the models as provided in R does not include an estimation of *p*-values. We therefore provide the *t*-values and assume that values of $t > 2$ indicate significant results. As can be seen below there are no values that fall into a range of *t*-values that could be considered ambiguous with regard to its statistical significance.

Results confirmed the use of spectral as well as durational vowel-internal cues in Experiment 2a ($b_{intercept} = -0.09$, $SE = 0.36$, $t = -0.25$; $b_{spectrum} = -0.778$, $SE = 0.17$, $t = -4.67$; $b_{rate} = -1.047$, $SE = 0.22$, $t = -4.78$; $b_{spectrum \times rate} = -0.297$, $SE = 0.34$, $t = -0.88$). Listeners also used the spectral and temporal information of the manipulated context to interpret the ambiguous vowels in Experiment 2b ($b_{intercept} = 0.61$, $SE = 0.35$, $t = 1.76$; $b_{spectrum} = 0.94$, $SE = 0.22$, $t = 4.24$; $b_{rate} = 0.78$, $SE = 0.17$, $t = 4.52$; $b_{spectrum \times rate} = -0.11$, $SE = 0.37$, $t = -0.31$). A baseline time window from 200 ms before vowel onset until 200 after vowel onset further confirmed that listeners did not anticipate their responses on basis of the contexts. No effect of either Spectrum or Rate was found for either the vowel condition (Experiment 2a: $b_{intercept} = -0.08$, $SE = 0.13$, $t = -0.64$; $b_{spectrum} = 0.08$, $SE = 0.22$, $t = 0.36$; $b_{rate} = 0.13$, $SE = 0.28$, $t = 0.48$; $b_{spectrum \times rate} = -0.17$, $SE = 0.48$, $t = -0.36$) or the context condition (Experiment 2b: $b_{intercept} = -0.13$, $SE = 0.15$, $t = -0.85$; $b_{spectrum} = 0.23$, $SE = 0.29$, $t = 0.77$; $b_{rate} = -0.06$, $SE = 0.21$, $t = -0.29$; $b_{spectrum \times rate} = 0.29$, $SE = 0.48$, $t = 0.60$).

3.2.2.2. Timing of the effects. To further investigate whether listeners used cues in the vowel earlier than cues in the context, and to test whether one type of cue (i.e., Spectrum or Rate) is taken into account earlier than the other, time points were calculated at which the different effects reached certain percentages of their maxima (McMurray et al., 2008; Miller et al., 1998; Ulrich & Miller, 2001). We thereby followed a method that was first suggested for the analysis of the timing in event related potentials (Miller et al., 1998) but has also been applied to the analysis of the uptake of acoustic cues as measured in eye tracking (McMurray et al., 2008). For example, we calculated at what point in time the effects of spectrum and rate reached 10% of their maxima. This was done for the vowel-conditions in Experiment 2a and for the context-condition in Experiment 2b. Note that the use of a measure relative to the maxima of the different effects rather than absolute values of target preferences at specific points in time allows for a comparison across different types of cues. To further explore whether possible differences in effects evolved over time, we calculated such time points at every 10% step of the maxima until 80% of the maxima was reached. The zero reference was chosen at 200 ms after vowel onset. 200 ms is the earliest point in time at which an influence of the acoustic cues on fixation behavior could possibly be observed. The time window within which the maximum values were defined was limited to 820 ms after vowel onset, matching the time window of the overall analyses and thus relating to the onset of the following context.

To compute the effects of Spectrum and Rate over time, we used the fixation preferences on the /ɑ/-words in the different spectrum and rate conditions on every sample recorded by the eye tracker (i.e., in 4 ms bins). Fixation proportions on the /ɑ/-words and /aː/-words were logistically transformed. Then the transformed proportions of fixations on the /aː/-words were subtracted from the transformed fixation proportions on /ɑ/-words to obtain a measure of fixation preference on /ɑ/ (as has been calculated for the overall analyses described in the preceding section). To further obtain measures of effects for spectrum and rate for the vowel-internal cues in Experiment 2a, the fixation preference on /ɑ/-words in the fast condition was subtracted from the fixation preference on /ɑ/-words in the slow condition. Similarly the fixation preference for /ɑ/-words in the high F_2 condition was subtracted from fixation preference on /ɑ/-words in the low F_2 condition. For Experiment 2b, the subtractions for the effects of spectrum and rate were reversed since the context-cues have a contrastive effect. These difference-measures were then smoothed by applying an 80 ms asymmetrical sawtooth window (cf. McMurray et al., 2008). This smoothing procedure eliminated ‘‘bumps’’ in the preference functions over time which may have occurred due to the small size of the time bins used to obtain a fine-grained timeline (i.e., we used every sample recorded by the eye tracker). The smoothing was applied such that the observations at each time point were the weighted sum of the observation of a range of values from 80 ms before up to a given time point. The weights of the preceding observations were so that the more distant the observation the less it contributed to the effects.

To statistically test differences between the points at which the effects of Spectrum and Rate reached the pre-specified percentages of their maximum values, the variance was estimated using a jackknife procedure (Ulrich & Miller, 2001). Time points for *N* subsets of the total participant population (*N*) were computed. Each subset contained data from *N*−1 participants; that is, each participant was excluded once. This method of using *N* sets of data from *N*−1 participants is well suited for the timecourse analysis of eye-tracking data. Eye-tracking data for single participants or single trials are often unreliable since the data can be sparse. Eye movements of a single participant are not continuous but fixations are separated by saccades. Only by averaging over a large number of trials and/or participants in each condition does a smooth increase of fixations towards the target objects emerge. Since averaged data, however, do not allow for statistical testing (statistical tests of group differences rely on the variances within and between groups) a jackknife-based method was used to assess timing differences in the uptake of spectral and temporal cues by means of statistical testing (see Miller et al., 1998; Ulrich & Miller, 2001, for a more detailed discussion on this issue).

Our jackknife-datasets of time points at which the effects of Spectrum and Rate reached certain percentages of their maxima in each of the two experiments were subjected to an ANOVA with Cue (spectrum vs. rate) as within-participant factor and Location (vowel vs. context) as between-participant factor. Importantly, *F*-values (and respective *p*-values) were adjusted for the fact that each participant contributed *N*−*1* times to the dataset (Ulrich & Miller, 2001). That is, *F*-values were divided by $(n-1)^2$ where *n* is the number of observations in each cell. Since the degrees of freedom of jackknife datasets are the same as for a respective simple dataset, *p*-values of the adjusted *F*-values could easily be assessed (Ulrich & Miller, 2001).

Fig. 6 shows how the effects of Spectrum and Rate in Experiment 2a and 2b increase up to their maxima between 200 ms and 820 ms after vowel onset. Table 3 reports the time points at which certain percentages of the maxima were reached as well as adjusted *F*- and *p*-values from the statistical analyses. Cue location (cues in the vowel vs. in the context) did not show any effect. Interestingly, however, the effects of Spectrum and Rate reached 40% and 50% of their maxima at different points in time. Adjacent percentage points (30%, 60%, and 70%) showed *p*-values between 0.1 and 0.2 just outside the ''marginally significant'' region. That is, as the effects of Spectrum and Rate build up, spectral cues appear to be used slightly earlier than rate cues, but this advantage decreases as the maxima of the effects are approached. No interaction between cue location and type of cue was found.

3.3. Discussion

Experiment 2a and 2b investigated the timecourse of the use of vowel-internal spectral and durational cues and compared this timecourse to the timecourse of the use of context information for the interpretation of ambiguous vowels. Overall analyses of click responses and eye-tracking

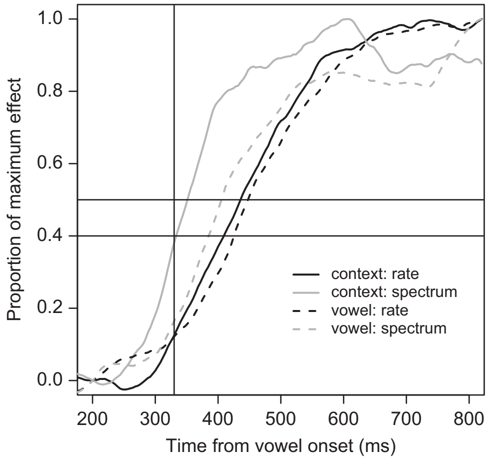

Fig. 6. Proportion of the maximum effects for the spectral and rate (duration) cues in the vowel vs. in the context. Proportions of the maxima are plotted over time from vowel onset shifted by 200 ms (i.e., the start of the displayed time window is 200 ms) until the end of the critical time window (see text for details). The vertical line indicates the average vowel offset shifted by 200 ms, the horizontal lines indicate the proportions of effects where the difference for spectrum vs. rate was significant (i.e., 40% and 50%).

Table 3
Results of the jackknife-based analysis of the uptake of different cues. ''Cross-points'' are the points in time (as measured from vowel onset) when the percentage points are reached for the different cues and cue locations. Times are given in milliseconds. *F*-values and *p*-values from the ANOVA have been adjusted for the repeated use of the data due to the jackknife procedure (see text for details).

	Mean cross-points vowel (ms)		Mean cross-points context (ms)		Cue (spectrum vs. rate)		Location (vowel vs. context)		Cue × location	
	Spectrum	Rate	Spectrum	Rate	F_c (1,62)	*p*	F_c (1,62)	*p*	F_c (1,62)	*p*
10%	319	325	287	331	0.28	.598	0.07	0.788	0.16	0.690
20%	350	376	311	360	1.49	0.226	0.69	0.410	0.14	0.709
30%	372	409	327	388	2.70	0.105	1.19	0.280	0.17	0.680
40%	393	433	342	417	4.37	<0.05	1.58	0.213	0.41	0.526
50%	413	456	359	444	4.70	<0.05	1.41	0.239	0.53	0.470
60%	440	486	376	474	2.73	0.104	0.81	0.372	0.36	0.550
70%	483	521	393	503	2.34	0.131	1.43	0.237	0.57	0.453
80%	528	567	426	542	0.93	0.338	0.71	0.403	0.23	0.634

data replicated the findings of Experiment 1. Listeners used spectral as well as durational/rate cues in the vowel for the identification of minimal pairs differing in the /ɑ/–/aː/ vowel contrast. Considering Fig. 5 as well as the regression weights for the different cues in the overall analyses, it appears that, first, the effects of Spectrum and Duration/Rate were about equal, and second, that the effort to roughly equate the effect size of the vowel vs. context manipulation was successful. Despite the minor differences in spectrum and duration for the vowel-internal cues, listeners were able to effectively exploit even these small differences. The version of the vowel that was only 12 ms shorter and had a second formant that was 41 Hz lower than the other version led to significantly more fixations and click responses on the word containing /ɑ/ than a longer version and a version with a high F_2. Similar effects (in the opposite direction) were found for the contextual manipulation in Experiment 2b. Listeners were able to use context information to interpret the ambiguous vowels. That is, listeners are sensitive to acoustic differences whether they occur in a vowel or are the result of perceived differences relative to a context.

In order to compare the timecourse of the uptake of different cues in the vowel against that of the context, time differences relative to the maxima of the effects were evaluated. This evaluation of relative effects allowed us to compare spectral and durational/rate cues that otherwise could not be compared as they represent different acoustic dimensions. Comparing differences in the uptake of vowel-internal cues vs. context information, no statistical difference could be found. This suggests that context information is taken into account as early as the upcoming sound is being processed (see the section ‘‘General Discussion’’ for possible implications of this finding). As for the different types of cues, a small but reliable difference between spectral and durational/rate cues was found as the effects built up and reached about 40–50% of their maxima. Since no difference was found between the uptake of vowel-internal cues and the use of context information our best estimate of the numerical difference between spectral and rate information is around 58 ms at the 40% crossover point and 64 ms at the 50% crossover point (40 and 43 ms for the vowel-internal differences, and 75 and 85 ms for the contextual differences, respectively). Note that this difference is less than or just about half the size of the average duration of the vowel tokens in Experiment 2 (∼130 ms).

4. General discussion

The current study had two main goals. The first was to establish at what point in time context information is taken into account for the evaluation of spectral and temporal cues in vowel perception. On the one hand, phoneme perception might always constitute the interpretation of speech cues relative to acoustic context. On the other hand, phoneme perception might first be based on local speech cues and contextual influences only arise at a later stage in processing. The second goal was to explore whether spectral and temporal cues are taken into account simultaneously or whether temporal cues are processed more slowly, since by definition they are spread out over time.

These questions were addressed in two experiments. In all experiments listeners categorized monosyllabic words as one of two members of minimal pairs such as /tɑk/–/taːk/. The first experiment was a simple categorization experiment. In a first part (Experiment 1a) the vowel of the stimulus words were taken from a stimulus grid that was created by orthogonally varying both vowel duration and F_2 (two of the most important cues to the Dutch /ɑ/–/aː/ contrast). The results of this experiment confirmed that Dutch listeners use both duration and F_2 to distinguish this vowel contrast. They further allowed us to select a range of ambiguous vowels to test the influence of context on vowel perception. In the second part (Experiment 1b) a new group of listeners categorized these ambiguous vowels, presented in manipulated context sentences. These context sentences had also undergone (orthogonal) manipulations of both speaking rate (the duration of all segment in the sentence were shortened or lengthened) and F_2 (the F_2 value of each of the vowels was increased or decreased). These manipulations to the context sentence were carried out independently of the target vowels. The results of this experiment showed that listeners' perception of the target vowels was significantly influenced by the auditory properties of the context sentences. These effects were contrastive. Ambiguous vowels were more often categorized as /ɑ/ (the short vowel with a low F_2) when the context was spoken slowly (i.e., had a longer duration) than when it was spoken fast. Moreover, more /ɑ/-responses were given when the context sentence had a generally high F_2 than when it had a low F_2. The results of Experiment 1 thus revealed robust, contrastive influences of context on the perception of temporal and spectral speech cues.

The second experiment was set up to provide more detailed information on cue uptake over time during the perception of the target vowels. To this end, listeners again categorized target words of minimal pairs such as /tɑk/–/taːk/. However, now in addition to the ‘‘click’’ responses their eye fixations on the printed words were measured. Since eye movements are sensitive to the processing of the auditorily presented input (with a delay of about 200 ms; e.g., Allopenna et al., 1998) this method allowed us to track listeners' recognition of the words over time, and thereby tap into the earliest moments at which the acoustic cues under investigation affected the vowel perception.

In Experiment 2a participants heard neutral sentences (there was no variation among trials in the duration or the general level of F_2 for the context sentences) which contained target words. The vowels in the target words were orthogonally varied across the trials with respect to both their duration and F_2 values (i.e., there were 4 target vowel conditions: low F_2 and short; high F_2 and short; low F_2 and long; high F_2 and long). In Experiment 2b participants heard ambiguous vowels on each trial, but these ambiguous vowels were presented in sentences that were varied with respect to their overall duration and F_2 value resulting in 4 context conditions (low F_2 and fast; high F_2 and fast; low F_2 and slow; high F_2 and slow). The combination of Experiments 2a and 2b allowed us to investigate whether contextually induced changes in cue perception on a target vowel arise at the same point in time as the perceptual consequences of vowel-internal differences in cue values. The results showed that contextual information influenced the processing of vowel-internal cues immediately. This was the case in both the spectral and in the temporal domain.

A significant precedence of vowel-internal cues over the use of contextual information would have provided strong evidence against a low-level, general auditory implementation of context effects. This difference, however, was not observed. Context information was used very rapidly to influence how upcoming sounds are perceived and interpreted. These findings provide support for previous arguments advocating the idea that context effects arise at early stages of word processing (Green et al., 1994, 1997; Lotto & Kluender, 1998; Newman & Sawusch, 2009; Sawusch & Newman, 2000; Watkins, 1991).

The current findings stress the fact that listeners are highly reliant on contextual information and are able to use context information as soon as it can provide useful adjustments to online perception. It should also be noted that listeners only used contextual information to adjust perception of the actual acoustic target information. For example, context information could, in principle, have been used by listeners to predict which of the two response options was going to be the more likely one. In the course of the experiment they could have learned that when the context is slow and has a specific timbre (due to high F_2) then the upcoming word is most likely the one with the /ɑ/-vowel – because this is what they had perceived on the first few trials. In this scenario, context effects could have preceded effects in the vowel-internal condition. Analyses of the baseline time windows, however, did not confirm this suggestion.[2] Rather, listeners used context information to rapidly adjust interpretation of the incoming ambiguous vowels.

Recently, neurophysiological evidence has also suggested an early locus of contextual influences on spectral cue perception. Sjerps, Mitterer, and McQueen (2011b) showed that the effect of spectral context effects can be observed already during the time window of the ERP component referred to as N1 (80–160 ms after target onset). In speech perception, the N1 time window is associated with pre-categorical levels of processing which clearly can be distinguished from later, phoneme-sensitive components such as the Mismatch Negativity (MMN; e.g. Sharma & Dorman, 2000). This suggests that the context effects found in the N1 time window take place at a stage of processing that is for a large part independent of linguistic exposure and hence occurs at early levels of processing.

While electroencephalographic information on early components can provide insight into the neural processes that precede behavior, they do not reveal to what extent these processes actually do influence behavior. The present eye-tracking study revealed that the influence of context information indeed impacts overt behavior at about the same time as the influence of sound-internal cues can be observed. As such, the current data suggest that early compensatory mechanisms play an important role in speech perception. This is not to say, however, that these findings show that higher level influences such as a mental-frame-of-reference based operation does not take place at all. The data by Johnson et al. (1999), described in the introduction, strongly argue that non-auditory based processes do play a role in speech perception to some extent. The current findings, however, suggest that in compensation for spectral and temporal properties of context sentences, early processes can account for a large part of the perceptual shifts that have been observed (e.g., Broadbent & Ladefoged, 1960; Sjerps et al., 2011a; Watkins, 1991).

As for the second question, regarding the uptake of spectral vs. durational cues, a small but reliable difference was found: the effect of spectral cues preceded the effect of rate. This suggests that the information value of durational cues is placed more heavily on the last part of the vowel than that of spectral information. The difference in the uptake of cues was significant at both 40% and 50% of the maxima of the effects. The fact that the precedence of spectral effects over duration/rate effects was not more pronounced (as one could have expected) is likely to result from the fact that the perceptual differences were relatively minor. First, in order to allow for a fair comparison between the timecourse of the context effect and the effect of vowel-internal cues (which was the main focus of the study), the current study employed an acoustic difference in the vowel categories of only 12 ms and 41 Hz (for the durational and spectral dimensions respectively). Our spectral and durational differences in the vowel were therefore much smaller than they can be expected to be in a full corpus of Dutch /ɑ/–/aː/ vowel contrasts (the average difference between /ɑ/ and /aː/ in our recordings was around 90 ms for duration and around 200 Hz for F_2). Importantly, however, unlike previous studies that investigated the uptake of multiple cues, the present study used stimuli in which the cues came available on the same segment of a word (i.e., in the vowel). Because the vowels were around 100 ms long this only provided a very restricted time window to observe any differences at all.

Table 3 shows that the most reliable difference between the information value of durational and spectral differences was reached at the 40% and 50% crossover points. These points were reached, however, only at vowel offset or slightly thereafter. This pattern indicates that when listeners perceive highly ambiguous vowels, the onset of the following consonant provided more information for the disambiguation of durational cues than for spectral cues. It is likely that the onset of the following consonant provides the definitive evidence about the duration of the preceding vowel, while the onset of the following consonant contains much less additional information about the spectral properties of a vowel. Overall, it seems that in the early parts of the vowel, information value of spectral and rate information are similar, but as the effects build up over time, spectral cues gain an advantage over durational cues. As both effects approach their maxima at the start of the following consonant, however, the statistical difference disappeared again.

To conclude, we observed that for the Dutch /ɑ/–/aː/ distinction, the informational value of temporal and spectral cues are distributed differently across the vowel. Spectral cues were used slightly earlier than durational cues in the perception of vowels. Importantly, the present results show that when listeners process temporal or spectral speech cues in the vowel, their perception of these cues is immediately influenced by acoustic properties of context sentences. These results argue for an important role of general auditory compensation processes in speech perception.

Acknowledgments

Both authors contributed equally to the experiments and to writing the paper.

We thank James M. McQueen and Holger Mitterer for advice during various stages of the project and three anonymous reviewers for helpful comments on the manuscript. Thanks go to Jelmer Wolterink, Thomas Kuijpers and Michael Wiechers for help with the experiments. Parts of this research have been presented at the First International Conference on Cognitive Hearing Sciences for Communication, Linköping, Sweden (2011).

Appendix A

Minimal word pairs selected for the eye-tracking experiments sorted by place of articulation of the preceding consonant. Note that in Dutch word-final devoicing applies. Words indexed by ⁎ were also used in Experiments 1 and 2.

Alveolar

daad–dat (‘‘deed’’–‘‘that’’), daag–dag (‘‘dare’’–‘‘day’’), knaak–knak (‘‘two euros and fifty cents’’–‘‘snap’’), knaap–knap (‘‘guy’’–‘‘handsome’’), staaf–staf (‘‘bar’’–‘‘staff’’), staand–stand (‘‘standing’’–‘‘posture’’), staat–stad (‘‘state’’–‘‘city’’), taak–tak (‘‘task’’–‘‘branch’’), zaad–zat (‘‘seed’’–‘‘sat’’), zaak–zak⁎ (‘‘shop’’–‘‘bag’’)

[2] Moreover, adding block or trial number as factors in the baseline model did not result in a main effect ($p > 0.4$) or an interaction with spectrum or rate ($p\text{s} > 0.36$). That is, listeners did not learn in the course of the experiment that context information rendered one answer more likely than the other.

labial

baan–ban (‘‘job’’–‘‘spell’’), baas–bas (‘‘boss’’–‘‘base’’), baat–bad (‘‘benefit’’–‘‘bath’’), maaɡd–macht⁎ (‘‘virgin’’–‘‘power’’), maak–mak (‘‘make’’–‘‘tame’’), maan–man (‘‘moon’’–‘‘man’’), maand–mand (‘‘month’’–‘‘basket’’), maat–mat (‘‘size’’–‘‘matt’’), smaak–smak (‘‘taste’’–‘‘smack’’), spaan–span (‘‘chip of wood’’–‘‘team’’, as in a team of horses), vaak–vak (‘‘often’’–‘‘section’’), vaat–vat (‘‘dishes’’–‘‘barrel’’), waak–wak (‘‘watch’’ as in watch-dog–‘‘hole’’ as in an ice-hole), waas–was (‘‘haze’’–‘‘laundry’’)

velar/glottal

ɡaas–ɡas⁎ (‘‘gauze’’–‘‘gas’’), haak–hak (‘‘hook’’–‘‘heel’’), kaas–kas (‘‘cheese’’–‘‘greenhouse’’), schaap–schap (‘‘sheep’’–‘‘shelf’’)

Appendix B

Based on the results of Experiment 1b, durational and spectral values for manipulation were determined. These manipulations were expected to shift the vowel perception in Experiment 2a (vowel-internal cues manipulated) by roughly the same amount as the contexts in Experiment 2b (only context manipulated). Table B1 displays the overall proportions of /ɑ/-responses in Experiment 1b for the different vowel steps. Using these values the perceptual effect of a speech cue could be calculated. The ‘‘Mean difference’’ column for spectral differences indicates that per step of 37.5 Hz listeners’ categorizations shift by a proportion of 0.13. This means that a shift of 1 Hz causes a shift in proportion /ɑ/-responses of (0.13/37.5 = 0.0035) 0.0035. For rate a similar calculation can be made. The ‘‘Mean difference’’ column for rate differences shows that per step of 8 ms, listeners’ categorization shifts by a proportion of 0.14. This means that a shift of 1 ms causes a shift in proportion /ɑ/ responses of (0.14/8.5 = 0.0166) 0.0166. Given these proportions we now could define a range of values that mimic the effect size obtained in Experiment 1b. In order to mimic the effect sizes induced by the context sentences spectral cues thus needed to span a range of 40.8 Hz around the ambiguous sound (0.140/0.0035 = 40), and duration needed to span a range of 12 ms (0.200/0.0166 = 12).

Table B1
Proportion of /ɑ/-responses in Experiment 1b to the vowel continua of duration and F_2 frequency, averaged over transition-types. Vowel durations (approximate) for the different steps are presented in milliseconds for the vowel part. F_2 (approximate) values are given in Hz. Values are approximate because they differ over the transition-types, see Table 1. ‘‘Spec. Means’’ indicate the row means for each spectral step-value. ‘‘Dur. Means’’ indicate the column means for the three different duration values. For both, the ‘‘Difference’’, reflects the numerical difference between two adjacent rows/columns. The ‘‘Mean Difference’’ reflects the average over the two ‘‘Difference’’ values.

F_2 values	Vowel duration					
	142	133	125	Spec. means	Difference	Mean difference
1220	0.72	0.61	0.43	0.59		
1258	0.6	0.42	0.28	0.43	0.15	
1295	0.44	0.31	0.21	0.32	0.11	0.13
Dur. means	0.59	0.45	0.31			
Difference	0.14	0.14				
Mean difference	0.14					

References

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America*, *116*, 1729–1738.

Adank, P., van Hout, R., & van de Velde, H. (2007). An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America*, *121*, 1130–1141.

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.

Baayen, H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effect modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J. (2008). Analyzing ’visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457–474.

Bates, D. M., & Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes* (version 0.999375-27) [software application].

Boersma, P., & Weenink, D. (2009). *PRAAT, doing phonetics by computer* (version 5.1) [computer program].

Brady, S. A., & Darwin, C. J. (1978). Range effects in perception of voicing. *Journal of the Acoustical Society of America*, *63*, 1556–1558.

Broadbent, D. E., & Ladefoged, P. (1960). Vowel judgments and adaption level. *Proceedings of the Royal Society*, *151*, 384–399.

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.

Crystal, T. H., & House, A. S. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America*, *72*, 705–716.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, *83*, 1553–1573.

Cutler, A., & Chen, H. C. (1997). Lexical tone in Cantonese spoken word-processing. *Perception & Psychophysics*, *59*, 165–179.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*, 507–534.

Escudero, P., Benders, T., & Lipski, S. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, *37*, 452–466.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children*. Doctoral Dissertation. Utrecht University. ISBN:90-76912-08-4.

Green, K. P., Stevens, E. B., & Kuhl, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics*, *55*, 249–260.

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics*, *59*, 675–692.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*, 305–312.

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America*, *120*, 2801–2817.

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*, 156–169.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic, and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*, 460–482.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Johnson, K. (2005). Speaker normalization in speech perception. In: D. B. Pisoni, & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Oxford, UK: Blackwell.

Johnson, K., Strand, E. A., & D′Imperio, M. (1999). Auditory visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359–384.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*, 59–69.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 736–748.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*, 602–619.

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *Journal of the Acoustical Society of America*, *113*, 53–56.

Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language*, *62*, 407–420.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccadic overhead. *Perception & Psychophysics*, *53*, 372–380.

McMurray, B., Clayards, M. A., Tanenhaus, M., & Aslin, R. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, *15*, 1064–1071.

McQueen, J. M., & Viebahn, M. (2007). Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology*, *60*, 661–671.

Miller, J. O., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology*, *35*, 99–115.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In: P. D. Eimas, & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ, USA: Erlbaum Associates.

Miller, J. L. (1987). Rate-dependent processing in speech perception. In: A. W. Ellis (Ed.), *Progress in the psychology of language*, Vol. 3 (pp. 119–157). London, UK: Erlbaum Associates.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088–2113.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *80*, 1297.

Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics*, *37*, 46–65.

Nooteboom, S. G., & Coden, A. (1984). *Het proces van spreken en verstaan, een nieuwe inleidingin de experimentele fonetiek*. Assen, The Netherlands: Van Gorcum.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

Quené, H., & Van Den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011a). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 978–996.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011b). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech*, *54*, 147–165.

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception & Psychophysics*, *62*, 285–300.

Sawusch, J. R., & Nusbaum, H. C. (1979). Contextual effects in vowel perception 1 anchor-induced contrast effects. *Perception & Psychophysics*, *25*, 292–302.

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics*, *68*, 1–16.

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America*, *107*, 2697–2703.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011a). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception, & Psychophysics*, *73*, 1195–1215.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011b). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia*, *49*, 3831–3846.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1074–1095.

Summerfield, Q., Haggard, M., Foster, J., & Gray, S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. *Perception & Psychophsyics*, *35*, 203–213.

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.

Ulrich, R., & Miller, J. O. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology*, *38*, 816–827.

van Heuven, V. J., van Houten, J. E., & De Vries, J. W. (1986). De perceptie van Nederlandse klinkers door Turken. *Spektator*, *15*, 225–238.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral envelope distortion. *Journal of the Acoustical Society of America*, *90*, 2942–2955.

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *96*, 1263–1282.

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *99*, 3749–3757.

West, P. (1999). Perception of distributed coarticulatory properties of English /l/ and /r/. *Journal of Phonetics*, *27*, 405–426.

Wilson, J. P. (1970). An auditory after-image in frequency analysis and periodicity detection in hearing. In: R. Plomp, & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing*. Leiden, The Netherlands: Sijthoff.