Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 27–38.

PHONEME RECOGNITION AS A FUNCTION OF TASK AND CONTEXT•

R.J.J.H. van Son and Louis C.W. Pols
Rob.van.Son@hum.uva.nl

Abstract

Phoneme recognition can mean two things, conscious *phoneme-naming* and pre-conscious *phone-categorization*. Phoneme naming is based on a (learned) label in the mental lexicon. Tasks requiring phoneme awareness will therefore exhibit all the features of retrieving lexical items. Phone categorization is a hypothetical pre-lexical and pre-conscious process that forms the basis of both word recognition and phoneme naming. Evidence from the literature indicates that phone-categorization can be described within a pattern-matching framework with weak links between acoustic cues and phoneme categories. This is illustrated with two experiments. The current evidence favors a lax-phoneme theory, in which all phonemic categories supported by the acoustic evidence and phonemic context are available to access the lexicon. However, the current evidence only supports segment-sized categories. It is inconclusive as to whether these categories are the same in number and content as the phonemes.

1 Introduction

In the phonetic literature, *phoneme recognition*, is generally used in two distinct senses. In one sense, phoneme recognition refers to “phoneme awareness”. In the other sense, it refers to a hypothetical intermediate, symbolic, representation in speech recognition.

Phoneme awareness comes with (alphabetic) literacy. Most literate people can easily identify or monitor phonemes in speech. This awareness builds on a “deeper”, automatic categorization of sounds into classes of which the listeners are *not* consciously aware. In the first sense, phoneme recognition is a form of word-recognition. In the second sense, it is a form of data-reduction that is hidden from awareness. We will refer to phoneme recognition in the former, conscious sense as *phoneme-naming* and in the latter, pre-conscious, sense as *phone-categorization*.

Phoneme-naming and phone-categorization are not identical. It is clear that in conscious phoneme-naming, labels are attached to the categories found in preconscious phone-categorization. However, the phoneme labels can also be obtained by recognizing the whole word first and then extracting the constituent phonemes from the lexicon (Norris et al. 2000). The conscious, lexical aspects of phoneme-naming will induce task effects in all experiments that rely on it. Obvious differences with phone-categorization are that lexical decisions are known to be competitive (winner-takes-all), frequency dependent, and prime-able. However, there

• First published in the Proceedings of the workshop on Speech Recognition as Pattern Classification, Nijmegen, Netherlands, July 11-13 2001, 25-30.

is no reason to assume that the underlying categorization is competitive, the frequency effects reported are intricate at best (McQueen and Pitt, 1996), and prime-ability might even be detrimental to word recognition. Furthermore, conscious awareness of phonemes and the associated attention allow the recruitment of “higher” mental modules that are inaccessible to unconscious processes (Baars, 1997).

2 The Units Of Speech

It is not clear whether the phoneme is really a “natural” element in recognition. Normally, phonemes are defined as distinctive feature bundles. That is, a phoneme is the smallest unit that will distinguish between words, e.g., [tEnt] versus [dEnt] or [kEnt]. In these examples, /t d k/ are phonemes that differ in the feature voicing (/t/-/d/), place of articulation (/t/-/k/), or both (/d/-/k/). Not all combinations of feature values that theoretically could be combined in a phoneme actually occur in a language. Languages ensure that differences between phonemes are large enough to be kept apart easily in both articulation and identification (Boersma, 1998; Schwartz et al., 1997). Of the 600 or more sounds that can be distinguished in the worlds languages, English uses less than 50. Furthermore, between languages, features and phonemes, can be defined differently, making for even more differences. For instance, both English [tEnt] and [dEnt] are transcribed as Dutch [tEnt] whereas both Dutch [tEnt] and [dEnt] are transcribed as English [dEnt].

Not all phoneme combinations are possible. [tEnt] can legally be changed into [tEnd]. But a change to [tEnk] results in an invalid English word. To get a valid English word, we have to change the place of articulation of the whole cluster /nt/, e.g., [tENk] is a valid English word. This is because there is a "phonotactic" rule in English that "forbids" /nk/ clusters. All languages have such rules, but they are different for each language (e.g., [tEnd] is *not* a valid Dutch word).

The phonotactic, and phonological, rules are a second (syntactic) layer that have to be added to the phonemes to get a workable set. In a sense, the *phonemes* define legal feature *combinations* and *phonotactic rules* define legal feature *sequences*. Therefore, it should not come as a surprise that phonemes and phonotactics are complimentary in speech recognition. People have difficulty producing and perceiving both phonemes with invalid feature combinations as well as feature (phoneme) sequences that violate phonotactic rules (Cutler, 1997). Speech that violates the combinatory rules of the features in a language will generally be mapped to the nearest valid phoneme sequence. This is a problem in second language learning as many (most) students never succeed in completely mastering the new phonemes and phonotactic rules.

We can capture thinking on phoneme recognition in terms of two extreme positions, which few phoneticians will actually defend. At the one extreme, the *obligatory phoneme hypothesis* states that all speech is internally represented as a string of phonemes and speech recognition is done on this phoneme string. Whether we use phonemes or features in this hypothesis is actually immaterial as all legal feature collections can be rewritten as legal phoneme sequences and vice versa. However, note that current theories of word recognition do not need phonemes or features. Most models would just as well work on “normalized” sound traces.

This brings us to the other extreme, the *lax phoneme hypothesis*. In this lax phoneme hypothesis, recognizing speech is tracking acoustic events for evidence of underlying words (or pronounceable sounds). As in all track reading, the tracks will be incomplete and ambiguous. To be able to recognize $\sim 10^5$ words from an unlimited number of speakers, on-line with incomplete, noisy evidence, the words have to be coded in ways that allow normalization, error correction, sequential processing and, not least, allow reproduction, as all humans are able to repeat new words.

A simple way of coding words in a robust way is to correlate acoustic events for sequential order and co-occurrence. Essentially, this is a “context-sensitive” clustering analysis in which the speech stream is (partially) categorized in a normalization and data-reduction step. After this data-reduction step, the remaining information can be processed further to fit the requirements of the lexicon.

The lax phoneme hypothesis states that the phoneme inventory and phonotactics capture the physiological and statistical regularities of (a) language. These regularities are used during speech recognition for the normalization and regularization of utterances as a precursor for lexical access.

To summarize the obligatory and lax phoneme hypotheses. The obligatory hypothesis states that all words (utterances) are mentally represented as phoneme strings. The lax phoneme hypothesis states that phonemes and phonotactics are features of a data-reduction process that selects and organizes relevant information, but doesn’t force decisions on segmental identity. In the lax hypothesis, missing information is ignored and strict categorization is deferred if necessary. In the obligatory phoneme hypothesis, missing information has to be provided (invented) during a forced phoneme categorization.

3 What Makes A Phoneme

One central presupposition that many theories on phoneme recognition share is that each phoneme has a unified (and unique) canonical *target* to which a realization can be matched (see Coleman, 1998 for evidence that this is a *perceptual* target). In this view, phone-categorization and phoneme-naming use the same “labels”. However, many *phones* of a given *phoneme* do not overlap in any perceptual representation, e.g., pre- and postvocalic liquids, glides, and plosives (c.f., aspirated and unreleased allophones of voiceless plosives). Whereas other phonemes share the same phones (but in different contexts), e.g., long and short vowels. This can be most clearly seen when phonotactically defined allophones in one language are distinct phonemes in another (dark and light /l/ in English or Dutch are two separate phonemes in Catalan).

Very often, only the context of a phone allows one to select the intended phoneme label. That these complex collections of "context dependent" phones are genuine objects and not artifacts of a procrustean theory is clear from the fact that both speakers and listeners can seamlessly handle the rather complex transformations to “undo” reduction, coarticulation, and resyllabyfication (e.g., "hall-of-fame" as /hO-l@-fem/ or “wreck a nice beach” as /rE-k@-nAi-spitS/).

Divergent allophones of a phoneme do not have to share any perceptual properties. Their unity at the phoneme level could, in principle, be completely arbitrary. That the allophones of a phoneme almost always do share some fundamental properties can be explained from the fact that phoneme inventories and the associated phonological and phonotactic rules evolve along the lines of maximal communicative *efficiency* in both production and perception (Boersma, 1998; Schwartz et al., 1997). This will favor “simple” inventories and rules. Still, each language-community can “choose” freely what variation it does or does not permit (Boersma, 1998). Our proposition is that phonemes are not only characterized by some perceptual "canonical form", but that phonotactical constraints and phonological rules are an integral part of phoneme identity. *A phone is the realization of a phoneme only in a certain context.* This is well illustrated by the fact that contexts that violate phonotactics hamper phoneme recognition (Cutler, 1997).

The lax phoneme hypothesis might at first not seem to require labeling each phone with a "master" phoneme label. However, for lexical access, each phone has to be reevaluated to determine its proper place in the utterance. For instance, /hO-l@-fem/ must be resyllabified to /hOl Of fem/ to be recognized as a three word phrase. The identity of the pre- and post-vocal /l/ and /f/ sounds is not trivial. At some level, even a lax-phoneme model should facilitate this exchange of allophones.

4 The Acoustics Of Phonemes

The previous discussion is "phonological" in nature in that no references were made to the acoustics, articulation, or perception of speech sounds. Features and phonemes are symbolic entities that have to be linked to acoustic categories to be of any use in speech communication. Two classical approaches to the perceptual categorization problem can be distinguished. First, are the *static clustering theories*. These theories assume that each phoneme is a simple perceptual category. This category is defined as a unit cluster in some perceptual space. Some, rather complicated, transformation is performed on the speech signal after which the kernel (center) of each phoneme realization will map to a point inside the boundaries of the perceptual area designated for that phoneme. The best known example of this kind of approach is the Quantal theory of speech (Ohala, 1989).

The second type of approach is *dynamical*. It assumes that the dynamics of speech generate predictable deviations from the canonical target realizations. These deviations can be "undone" by the extrapolation of the appropriate parameter tracks (dynamic specification, see Van Son, 1993a, 1993b) or by some detailed modeling of the mechanical behavior of the articulators (Motor theory). Experimental evidence for any of these theories has been hotly disputed. As Nearey (1997) rightfully remarks: Proponents of both approaches make such a good case of disproving the other side, that we should believe them both and consider both disproved.

5 Experimental Illustration

An experiment we performed some years ago illustrates the problems of theories relying on static or dynamic specification (Pols and Van Son, 1993; Van Son, 1993a, 1993b). In our experiment we compared the responses of Dutch subjects to isolated synthetic vowel tokens with *curved* formant tracks (F_1 and F_2) with their responses to corresponding tokens with *stationary* (level) formant tracks. We also investigated the effects of presenting these vowel tokens in a synthetic context (/nVf/, /fVn/).

Nine formant "target" pairs (F_1, F_2) were defined using published values for Dutch vowels. These pairs corresponded approximately to the vowels /iuyIoEaAY/ and were tuned to give slightly ambiguous percepts. For these nine targets, smooth formant tracks were constructed for F_1 and F_2 that were either level or parabolic curves according to the following equation (see figure 1):

$F_n(t) = Target - \Delta F_n \cdot (4 \cdot (t/D)2 - 4 \cdot t/D + 1)$
in which:
$F_n(t)$: Value of formant n (i.e., F_1 or F_2) at time t.
ΔF_n: Excursion size, F_n(mid-point) - F_n(on/offset).
ΔF_1= 0, +225 or -225, ΔF_2= 0, +375 or -375 (Hz)
Target: Formant target frequency.
D: Total token duration (0 < t < D).

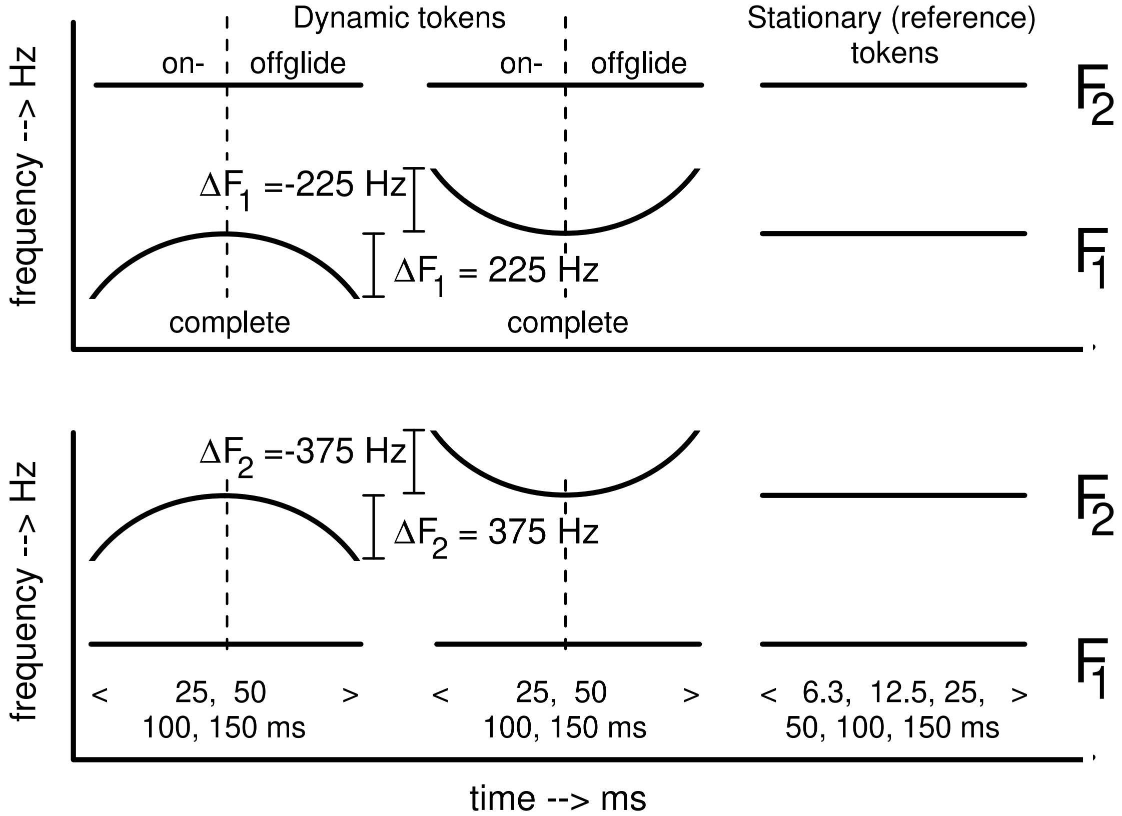

Figure 1. Formant track shapes as used in the experiments discussed in section 5. The dynamic tokens were synthesized with durations of 25, 50, 100, and 150 ms. The stationary tokens were synthesized with durations of 6.3, 12.5, 25, 50, 100, and 150 ms. The dynamic tokens were also synthesized as onglide- and offglide-only tokens, i.e., respectively the parts to the left and right of the dashed lines.

No tracks were constructed that would cross other formant tracks or F_0. All tracks were synthesized with durations of 25, 50, 100, and 150 ms (see for more details: Pols and Van Son, 1993; Van Son, 1993a, 1993b). Stationary tokens with level formant tracks (i.e., $\Delta F_1 = \Delta F_2 = 0$) were also synthesized with durations of 6.3 and 12.5 ms. Of the other tokens (with either ΔF_1= +/-225 Hz or ΔF_2= +/-375 Hz), the first and second half of the tracks, i.e., on- and offglide-only, were also synthesized with half the duration of the "parent" token (12.5, 25, 50, and 75 ms). Some other tokens with smaller excursion sizes were used too, these will not be discussed here (but see Pols and Van Son, 1993; Van Son, 1993a, 1993b). In experiment 1, tokens were presented in a pseudo-random order to 29 Dutch subjects who had to mark the orthographic symbol on an answering sheet with all 12 Dutch monophthongs (forced choice).

In experiment 2, a single realization each of 95 ms synthetic /n/ and /f/ sounds were used in mixed pseudo-syllabic stimuli. Static and dynamic vowel tokens from the first experiment with durations of 50 and 100 ms and mid-point formant frequencies corresponding to /I E A o/ were combined with these synthetic consonants in /nVf/ and /fVn/ pseudo-syllables. The corresponding vowel tokens with only the on- or off-glide part of parabolic formant tracks (50 ms durations only) were used in CV and VC structures respectively. For comparison, corresponding stationary vowel tokens with 50 ms duration were also used in CV and VC pseudo-syllables. Each vowel token, both in isolation and in these pseudo-syllables, was presented twice to 15 Dutch subjects who were asked to write down what they heard (open response).

The speech recognition theories discussed above make clear predictions about the behavior of our listeners. Static theories predict that vowel identity is largely unaffected by formant dynamics. Dynamic theories predict some compensation for

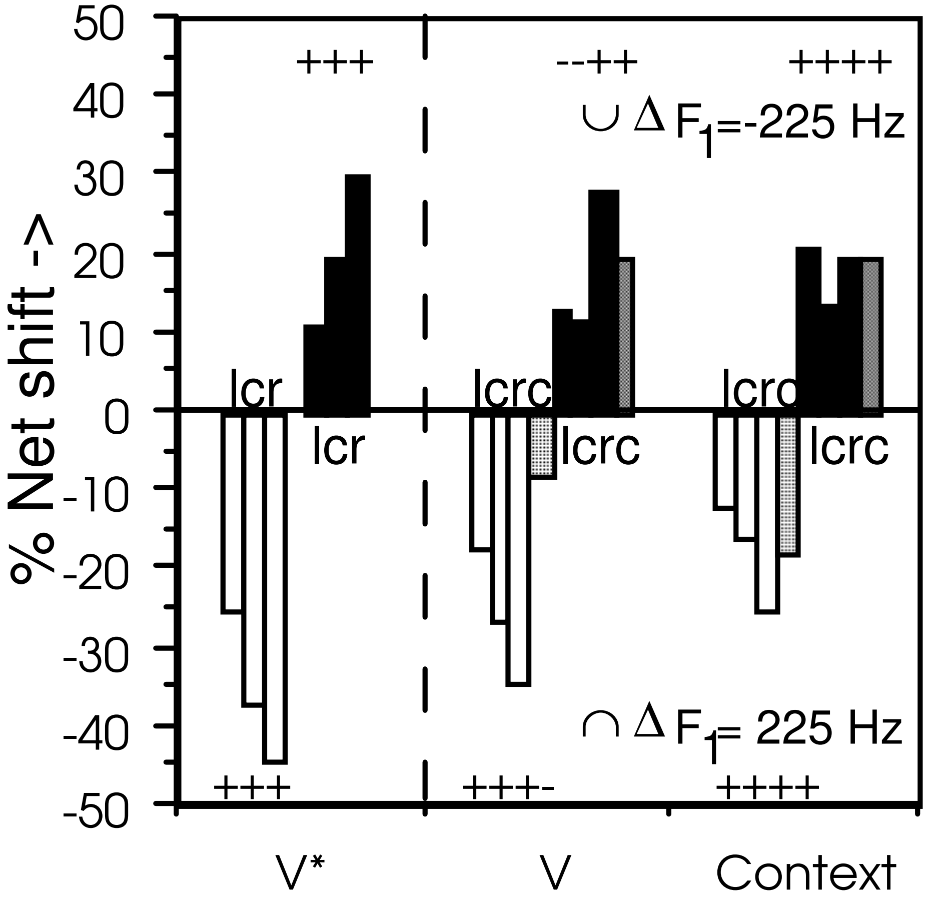

Figure 2. Net shift in responses as a result of curvature of the F_1. 'V*' are the results of the first experiment (all tokens pooled on duration, n >= 696). 'V' and 'Context' are the results of the second experiment with vowel tokens presented in isolation ('V' n=120, left; n=90, right), or in context, CV, CVC, VC; C one of /n f/ ('Context' n=240, left; n=180, right). Gray bars: 100 ms tokens, white/black bars: 50 ms tokens, l=onglide-only, c=complete, r=offglide-only tokens. +: significant ($p < 0.001$, sign test), -: not significant. Results for F_2 were comparable but weaker.

reduction in dynamic stimuli. In our case, all dynamic theories would predict formant track extrapolation in some form (perceptual overshoot see Pols and Van Son, 1993; Van Son, 1993a, 1993b).

For each response to a dynamic token, the position in formant space with respect to the static token was determined. For instance, an /E/ response to a dynamic token was considered to indicate a higher F_1 perception and a lower F_2 perception than an /I/ response to the corresponding static token. By subtracting the number of lower dynamic responses from the number of higher dynamic responses, we could get a *net-shift* due to the dynamic formant shape (testable with a sign test). Analysis of all material clearly showed a very simple pattern over all durations: Responses *averaged* over the trailing part of the F_1 tracks (figure 2). The same was found for the curved F_2 tracks (not shown), although here the effects were somewhat weaker and not always statistically significant (see Pols and Van Son, 1993; Van Son, 1993a, 1993b for details).

The use of vowels in /n/, /f/ context had no appreciable effect except for a decreased number of long vowel responses in open "syllables" (not shown). In accord with the Dutch phonotactical rule against short vowels in open syllables. However, this lack of effect could be an artifact from an unnatural quality of the pseudo-syllables.

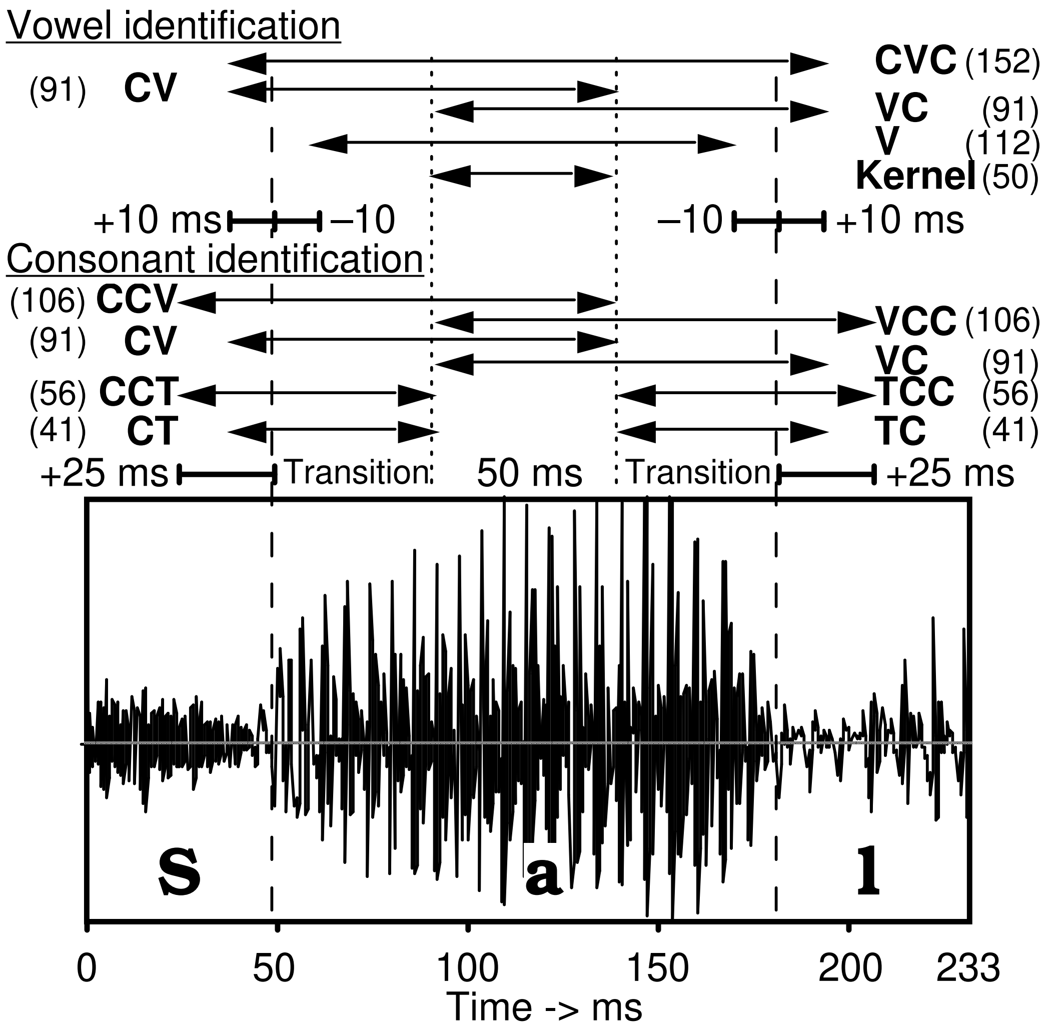

Figure 3. Construction of tokens from Consonant-Vowel-Consonant speech samples, taken from connected read speech, and their median durations (between brackets). Example for a /Sa:l/ speech sample. Vowel durations were always 100 ms or more (median: 132 ms). Scale lines marked with +/–10 ms and +25 ms are relative displacements with respect to the vowel boundaries (outer pair of dashed lines). Only the "vowel-transition" parts of the tokens were defined with variable durations (>= 25 ms, between the outer and inner pair of dashed lines). The Kernel part and both types of Consonant parts (short, C, and longer, CC,) of the tokens were defined with fixed durations (50, 25, and 10 ms, respectively).

Contrary to the predictions of the *dynamic* models of speech recognition, there was *no extrapolation* found. Contrary to the predictions of the *static clustering theories*, the *kernel* was not exclusively used for identification. None of the theories predicted, or can even explain, the prevalence of averaging responses to dynamic stimuli. No segment internal cues seem to be used to compensate for the natural variation in vowel acoustic. Therefore, we should look for contextual cues.

6 Pattern-Recognition Models Of Phoneme Recognition

In two papers, Nearey points out that most current theories on speech perception, like those discussed above, assume that there are strong links between the symbolic, i.e., phoneme, and either the articulatory or the perceptual level, or both (Nearey, 1992, 1997). He calls these theories *strong*. In contrast, Nearey puts forward the empirical hypothesis that '[S]*peech cues can be directly mapped onto phonological units of no* *larger than phoneme size*' (Nearey, 1997). Nearey states that both speakers and listeners are able to exploit almost any regularity in the correspondence between acoustics and phoneme identity. Hence, there are only weak links between the articulatory, perceptual, and symbolic levels, which merits the name *weak* theories.

What distinguishes strong and weak theories of perception is the importance of *local context* (Nearey, 1997). In an ideal *strong* theory, the acoustics of speech are primarily linked to local "features" of articulation or perception. Hence, ambiguities can be resolved within the phoneme itself and any distant, perisegmental, cues are redundant, i.e., they do not supply new information. Weak theories along the lines drawn by (Nearey, 1992, 1997), allow for any regularity to become distinctive (including visual ones). This implies that the relevant cues must be extracted and interpreted within a wider context, which does supply new information not available from within the phoneme boundaries. In a weak theory, multiple sources of information are integrated to come to a single phoneme percept. In the view of Nearey's account (Nearey, 1997), the "strength" of a perception model would scale with the scope of the speech window relevant to phoneme identification. This can be operationalized by plotting identification performance as a function of the amount of speech available, i.e., the distance to the "segment proper".

The *weak* approach to phoneme perception of (Nearey, 1992, 1997) fits in the pattern-recognition framework of (Smits, 1997). In this framework, all relevant acoustic "events" in the neighborhood of a segment are collected to decide on the presence of a phoneme. What is *actually* used to recognize a phoneme is determined by what (other) acoustic and visual cues or phonemes are present, or even by "circumstantial evidence".

Strong theories of speech perception predict that listeners will only use cues from within a segment. Recognition performance will reach a ceiling at the boundaries of the segments. Visual information, e.g., lip-reading, and the associated synchronization problems are generally ignored in these theories, although motor-theory based models could in principle integrate it. Adding speech from outside the phoneme boundaries will not improve identification. The pattern-matching framework of perception, e.g., the weak theory of (Nearey, 1992, 1997), predicts that phoneme identification will benefit from *any* "extra" speech, irrespective of its modality or relative position in the utterance.

The pattern-matching framework can be illustrated by (Smits, 2000) and by a study of our own (Van Son and Pols, 1999). In the latter study, we constructed gated tokens from 120 CVC speech fragments taken from a long text reading where the sentence accent on the vowel was noted (see Van Son and Pols, 1999 for details). The tokens were divided into vowel kernel (kernel, the central 50 ms), vowel transition (T, everything outside the kernel), a short consonant part (C, 10 ms), and an overlapping longer consonant part (CC, 25 ms). Gated tokens were constructed from these sub-segments according to figure 3. These tokens were randomized for vowel identification (Kernel, V, CV, VC, CVC tokens), pre-vocalic consonant identification (CT, CCT, CV, CCV tokens), and post-vocalic consonant identification (TC, TCC, VC, VCC tokens). Listeners were asked to identify vowels (17 subjects), and pre-, or post-vocal consonants (15 subjects for both) by picking the relevant orthographic symbol on a CRT screen. Subjects could pick any legal phoneme, except schwa as well as some unusual consonants (/JGg/, affricates). For more details, see (Van Son and Pols, 1999).

The results were analyzed in terms of the $\log_2$(response-perplexity) which is a measure of the missing information (in bits) and is measured on a ratio scale. Contrary to its complement, the mutual information or transmission rate, the missing

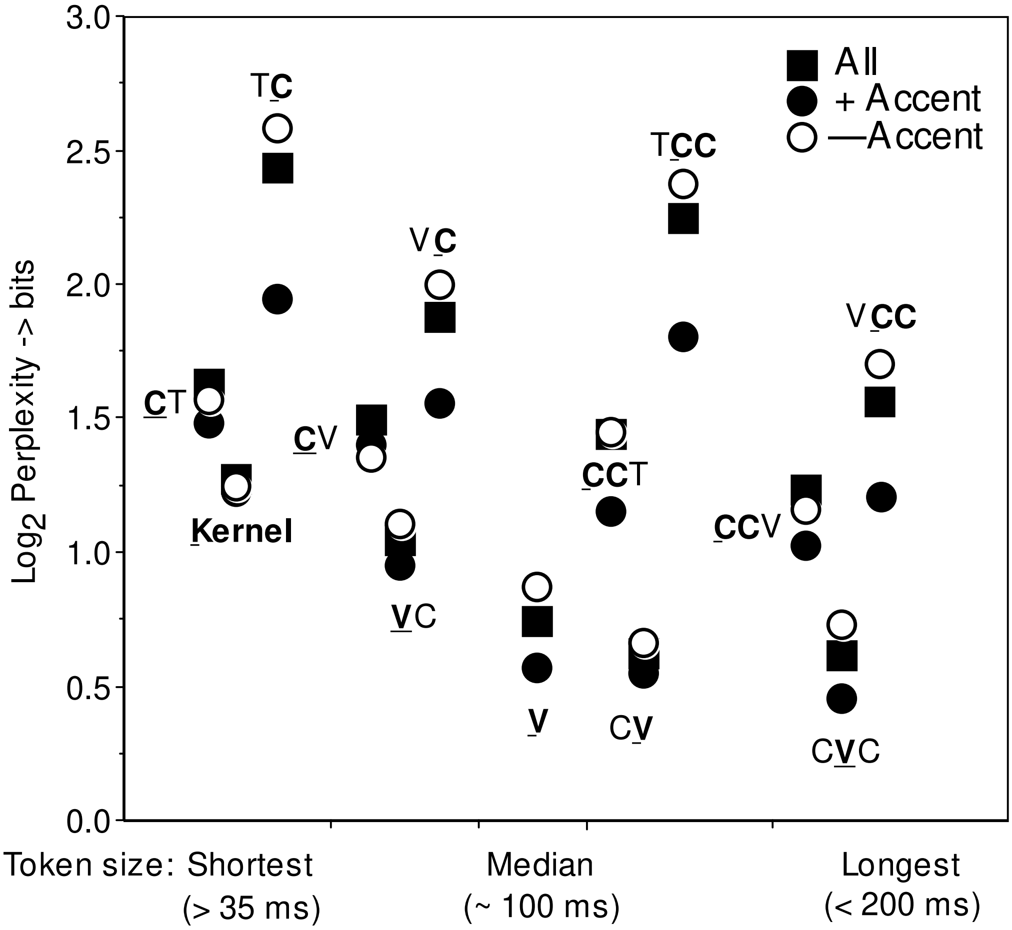

Figure 4. Log$_2$ perplexity values of *vowel* and *consonant* identification (in *pre*- and *post*-vocal position) for the individual stimulus token types. Given are the results for all tokens pooled ("All") as well as for vowels with and without sentence accent separately (+ and – Accent respectively). Long-short vowel and consonant voicing errors were ignored. Chance response levels would result in a log$_2$ perplexity of 2.93 bits for vowels and 3.83 bits for consonants. <u>Kernel</u>, <u>V</u>C, <u>V</u>, C<u>V</u>, C<u>V</u>C: Vowel identification (lower/center symbols). <u>C</u>T, <u>C</u>V, <u>CC</u>T, <u>CC</u>V: pre-vocal consonant identification (middle/left symbols). T<u>C</u>, V<u>C</u>, T<u>CC</u>, V<u>CC</u>: post-vocalic consonant identification (upper/right symbols). Statistical analysis of the error rates showed that all differences between *Accented* and *Unaccented* tokens were significant ($p < 0.01$) except for vowel identification in Kernel and VC tokens (Chi-square test). Furthermore, all differences between the *All* tokens were significant, except for vowel identification between CV and CVC tokens (McNemar's test).

information is insensitive to the size of the stimulus set. The results are summarized in Figure 4.

Two conclusions can be drawn from figure 4. First, phoneme identification benefits from extra speech, even if it originates from outside the segment proper, e.g., adding a vowel kernel to a transition-consonant fragment. Second, adding contextual speech in front of a phoneme improves identification more than appending it at the back. Our results could be best explained by assuming that, as predicted by the weak theory of speech perception, listeners use speech cues from far outside the segment to identify it. Furthermore, the prevalence of early (pre-pended) speech cues over late (appended) cues indicates that phoneme-identification (a *naming* task) is a fast process in which *label*-decisions are made as early as possible, disregarding subsequent cues.

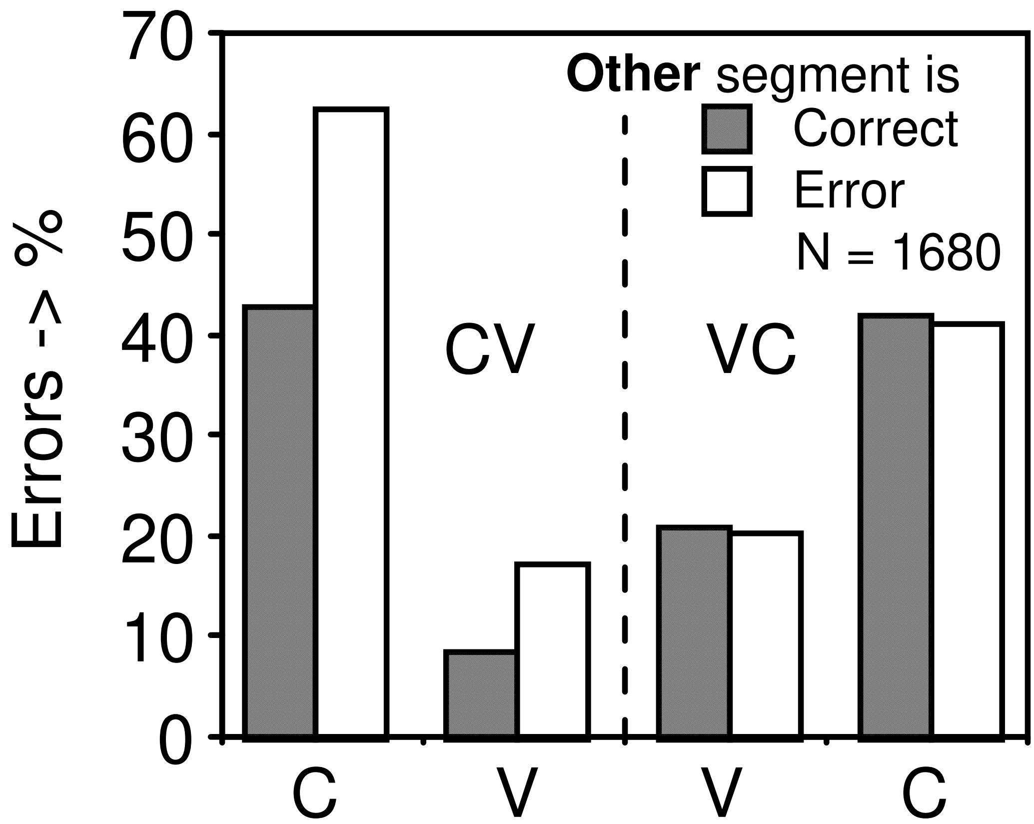

Figure 5. Error rates for vowel and consonant identification in CV- and VC-type tokens with respect to the correct and incorrect identification of the other segment in the same token. Voiced/Voiceless errors in consonants and Long/Short errors in vowels were ignored. Differences are statistically significant for the CV tokens only (Chi-square = 28.7, n = 1, p < 0.01).

7 Phonemes In Context

Phoneme production (articulation) is highly context dependent. Coarticulation is one of the prime sources of variation in phonemes. It changes all aspects of speech to such an extent that no genuine acoustic invariant has been found in 50 years of research. So it would be no surprise if phoneme recognition itself would be context dependent.

A reanalysis of classical studies on dynamic theories of phoneme recognition (Van Son, 1993a, 1993b), showed that all of these studies could be interpreted in terms of purely phonemic-context effects: Only if the appropriate context was identified, was there compensation for coarticulation. Furthermore, the *amount* of perceptual compensation depended *only* on the context, and was independent of the *size* of any dynamic aspect of the speech. In an extensive in-depth analysis of earlier experiments, Nearey gives very convincing arguments for the use of phoneme-sized, symbolic context in phoneme recognition (Nearey, 1990). If we summarize these studies, listeners seem to interpret acoustic cues not with respect to their *acoustic* context, but instead with respect to their *phonemic* context (Nearey, 1997; Pols and Van Son, 1993; Van Son, 1993a, 1993b; Nearey, 1990, 1992).

This phonemic-context effect can be illustrated with results from our own study (Figure 5, Van Son and Pols, 1999). Both vowels and consonants in CV tokens were identified better when the other member of the pair was identified correctly than when it was identified incorrectly. The fact that nothing was found for the VC tokens makes it less likely that this effect was only due to the fact that a better articulated vowel implicated a better articulated consonant. However, we do not have an explanation of the difference between CV and VC tokens.

8 A Synthesis?

It is probably too early to present a real synthesis of theories on phoneme recognition. However, several points can already be made. A good case has been made for a purely bottom up model (Norris et al. 2000), that fits a weak pattern-matching framework (Nearey, 1992, 1997; Smits, 1997). The input will be a combination of acoustic and visual "events" in the speech signal that map directly onto symbolic entities of phoneme size (Nearey, 1992, 1997; Smits, 1997). The output are phoneme-sized categories used to access the lexicon or articulation (Coleman, 1998).

What is still needed is a mechanism to normalize and combine speech cues into phoneme-sized categories that compensates for coarticulation and reduction. Most evidence is compatible with a view that the normalization and compensation depends on the *phonemic* (symbolic) context (Van Son, 1993a, 1993b; Van Son and Pols, 1999; Nearey, 1990).

A little is already known of the categorization process. First of all, acoustic cues are "recycled" and each individual cue can be used for more than one phoneme, e.g., vowel duration in vowel identification and plosive voicing (Nearey, 1990). Furthermore, when present, visual information is used in the categorization process, e.g., the McGurk effect (McGurk and MacDonald, 1976). Categorization is also lax in that ambiguities are preserved (Schouten and Van Hessen, 1992; Van Hessen and Schouten, 1992), as is demonstrated in the Ganong effect (Borsky et al., 1998), where categorical boundaries shift in response to lexical expectations, and phoneme restoration. It seems as if during categorization as much data reduction is performed as possible without discarding relevant cues.

This can be visualized as the construction of an ASR-like lattice of possible "phone(me)-categories" and their "activations" (or sublabeling, Van Hessen and Schouten , 1992). These activations build up over time (Schouten and Van Hessen, 1992; Van Hessen and Schouten , 1992). The lattice would contain any phoneme (or phone-category) that is supported by the acoustic, visual, and contextual evidence. That is, speech recognition would use a lax-phoneme hypothesis. This phone(me)-lattice is pre-conscious and can only be accessed by way of the lexicon or by articulatory recoding (Coleman, 1998). The fact that the phone-categories can directly be recoded for articulation, bypassing the lexicon, suggests that the categorization process itself is located outside the lexicon (c.f., Coleman, 1998).

Note that none of the evidence discussed here actually proves that the phone-categories are in fact phonemes. The only requirements are that these categories map to phoneme-*sized* chunks of speech and that they can be used to access phoneme labels in the lexicon and the articulatory apparatus for "echoing". This suggests that there could well be more phone-categories at this level than there are phoneme-labels in the lexicon. A guess would be some kind of allophones, but nothing definite is known about these categories or their exact nature.

There remains the question of how to investigate the phone-categorization process. The principal experiments into phoneme recognition rely on either lexical access of some sort, e.g., phoneme identification or monitoring, or the categorical perception paradigm. The latter task might use articulatory recoding (Coleman, 1998). Another way to probe the categorization process is to rely on articulatory recoding by using shadowing tasks. In shadowing experiments, the use of the lexicon can be manipulated (or blocked) to reveal the underlying processes. This is the direction we are currently investigating with some new experiments.

Acknowledgements

This work was made possible by grant nrs. 300-173-029 and 355-75-001 of the Netherlands Organization for Scientific Research (NWO).

References

Baars B.J. (1997). A thoroughly empirical approach to consciousness: Contrastive analysis, in Ned Block, Owen Flanagan, and Guven Guzeldere (eds.), *The nature of consciousness*, Cambridge MA: MIT/Bradford Books.

Boersma P. (1998). *Functional Phonology, formalizing the interactions between articulatory and perceptual drives*, PhD thesis University of Amsterdam.

Borsky S., Tuller B. and Shapiro L.P. (1998). How to milk a coat: The effects of semantic and acoustic information on phoneme categorization. *Journal of the Acoustical Society of America* 103, 2670-2676.

Coleman J. (1998). Cognitive reality and the phonological lexicon: A review. *Journal of Neurolinguistics* 11, 295-320.

Cutler A. (1997). The comparative perspective on spoken-language processing, *Speech Communication* 21, 3-15.

McGurk H. and MacDonald J. (1976). Hearing lips and seeing voices. *Nature*, Dec. 1976, 746-748.

McQueen J.M. and Pitt, M.A. (1996). Transitional probability and phoneme monitoring. *Proceedings of ICSLP'96*, Philadelphia, volume 4, 192-197.

Nearey T.M. (1990). The segment as a unit of speech perception. *Journal of Phonetics* 18, 347-373.

Nearey T.M. 1992. Context effects in a double-weak theory of speech perception. *Language and Speech* 35, 153-171.

Nearey T.M. 1997. Speech perception as pattern recognition. *Journal of the Acoustical Society of America* 101, 3241-3254.

Norris D., McQueen J.M., and Cutler A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* 23, 299-325.

Ohala, J.J. ed. (1989). Special issue on Quantal theory of speech, *Journal of Phonetics* 17, 1-157.

Pols L.C.W. and Van Son R.J.J.H. (1993). Acoustic and perception of dynamic vowel segments. *Speech Communication* 13, 135-147.

Schouten M.E.H. and Van Hessen A.J. (1992). Modeling phoneme perception. I: Categorical perception. *Journal of the Acoustical Society of America* 92, 1841-1855.

Schwartz J.-L., Boë L.-J., Vallée N. and Abry C. (1997). The dispersion-focalization theory of vowel systems, *Journal of Phonetics* 25, 255-286.

Smits R. (1997). A pattern recognition framework for research on phonetic perception. *Speech Hearing and Language Work in Progress* 9. Department of Phonetics and Linguistics, University College, London, 195-229.

Smits R. (2000). Temporal distribution of information for human consonant recognition in VCV utterances. *Journal of Phonetics* 27, 111-135.

Van Hessen A.J. and Schouten M.E.H. (1992). Modeling phoneme perception. II: A model of stop consonant discrimination. *Journal of the Acoustical Society of America* 92, 1856-1868.

Van Son R.J.J.H. (1993a) *Spectro-temporal features of vowel segments*. PhD thesis University of Amsterdam.

Van Son R.J.J.H. (1993b). Vowel perception: a closer look at the literature. *Proceedings of the Institute of Phonetic Sciences*, University of Amsterdam, 17, 33-64.

Van Son R.J.J.H. and Pols L.C.W. (1999). Perisegmental speech improves consonant and vowel identification. *Speech Communication* 29, 1-22.