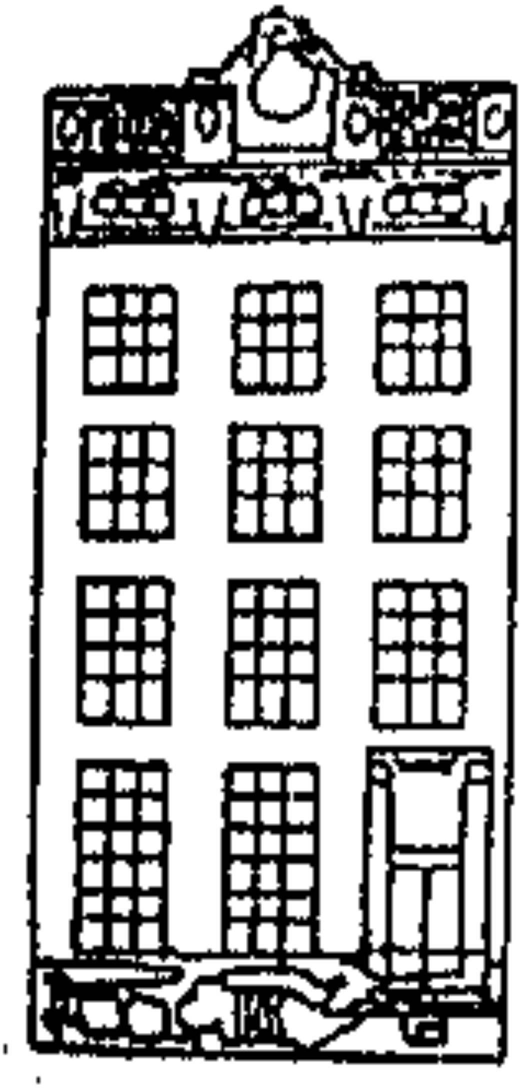
PHONEME RECOGNITION AS A FUNCTION OF TASK AND CONTEXT

R.J.J.H. van Son and Louis C.W. Pols
Institute of Phonetic Sciences IFA/ACLC
University of Amsterdam, The Netherlands
Rob.van.Son@hum.uva.nl



Introduction

Phoneme recognition has 2 meanings**:**

1 Phoneme naming

2 Phone categorization

Where *phone categorization* precedes *phoneme naming*

Ad 1: **Phoneme naming**

- Consious (identification)
- Lexical (results in a label)
- Competitive (winner takes all)
- Prime-able
- Frequency sensitive

Ad 2: **Phone categorization** (hypothetical)

- Pre-consious / 'On-line'
- Pre-lexical
- Many categories can be activated
- Unprime-able?
- Frequency effects are 'intricate'

Units of speech

Phones or *Phonemes* are considered here to be the *units* of speech (which is an over-simplification)

Definition of Phonemes:
Smallest "unit of difference" between words.
Phonemes are described as ***Feature bundles***

Examples:			(feature difference)
[tEnt]	<-->	[dEnt]	(voicing)
[tEnt]	<-->	[kEnt]	(place of articulation)
[dEnt]	<-->	[kEnt]	(both)

Not all possible feature bundles are legal phonemes:
> 600 phonemes known worldwide
English uses **< 50**

Every language differs in the way it defines features

Example: Voicing

English	/tEnt/ & /dEnt/	--> Dutch	[tEnt]
Dutch	/tEnt/ & /dEnt/	--> English	[dEnt]

Phonotactics

Not all phoneme combinations are legal

Phonotactic & phonological rules define legal phoneme and feature combinations
These rules define the smallest possible differences between words

Examples:

[tEnt]	<-->	[tEnd]	(English, **Dutch*)
[tEnt]	<-->	[tEnk]	(**English, *Dutch*)
[tEnt]	<-->	[tENK]	(English, Dutch)

Phonotactic & phonological rules are a syntactic layer over the phoneme sets.
Phoneme inventory and phonology are optimized with respect to each other.

Phonemes define legal feature *combinations.*
Phonological (phonotactical) rules define legal feature *sequences.*

Both "illegal" phonemes and "illegal" phoneme sequences hamper production and perception

The role of Phonemes in speech recognition

Two opposite (extreme) hypotheses:

A) *Obligatory phoneme hypothesis*
All speech is converted to a string of phoneme symbols before lexical access. Phoneme categorization is absolute and obligatory.

B) *Lax phoneme hypothesis*
Phonemes (or phones) are the result of prelexical regularization and data reduction processes that extract the relevant acoustic information.
The phone(me)s are clustering artefacts of the extracted information, i.e., they represent the prefered "acoustic events".
In this hypothesis, categorization is partial and can be defered.

What makes a phoneme?

Does every phoneme have a unified and unique canonical target?
(both in production and perception)

Unlikely cf.:
different phones/same phoneme
/l/ in [hOl] and [lOw]   (dark vs light)
/t/ in [dEnt] and [tEnd]   (unreleased vs aspirated)

same phones/different phonemes ("bad bet" exp)
/E/ vs/ae/ in   [bEd] and [baet] (short vs long)
/d/ vs /t/ in   [baed] and [bEt] (-/+ voiced)
/I/ vs /O/ in   [mIljun] and [bijOsko:p] (Dutch)

A *phone* is a realization of a *phoneme* only in a certain *context.*
Allophones of a phoneme do not have to have anything in common at all.

Proposition:
The identity of a phone in *context* is completely at the discretion of the language and how it optimizes the trade-off between ease of production and perception.

The acoustics of phonemes

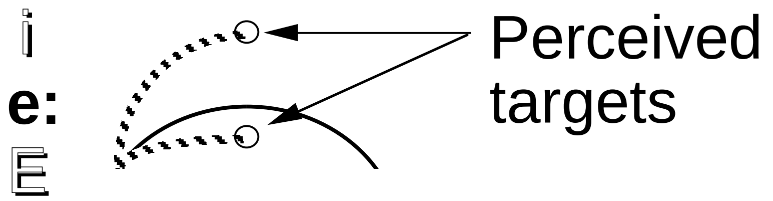
Classical approaches:

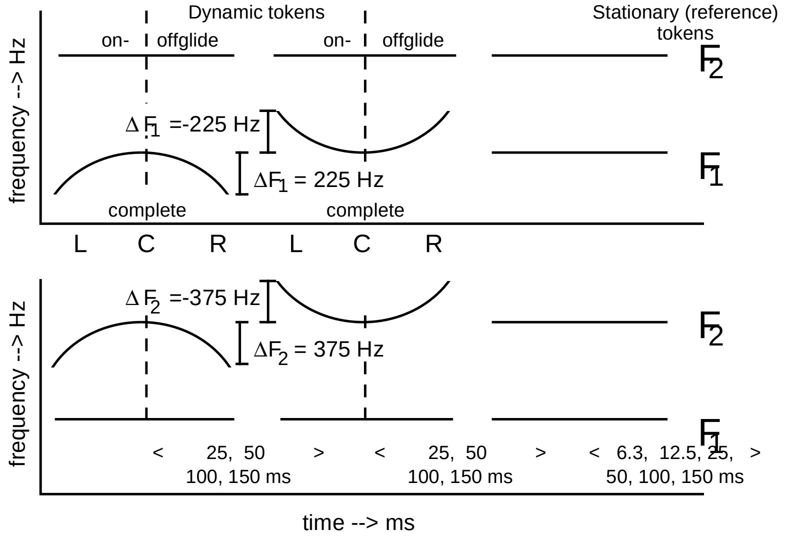
A) Static clustering theories
Each phoneme is a simple, continuous category in some perceptual space.
Requires rather complex acoustic transformations (normalizations).

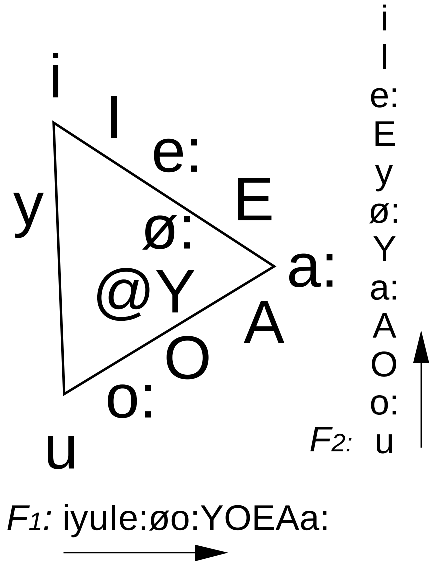
B) Dynamical specification
The dynamics of speech generates predictable deviations from the canonical target that can be undone by extrapolation of suitable parameter tracks or inverse modelling (motor theory).

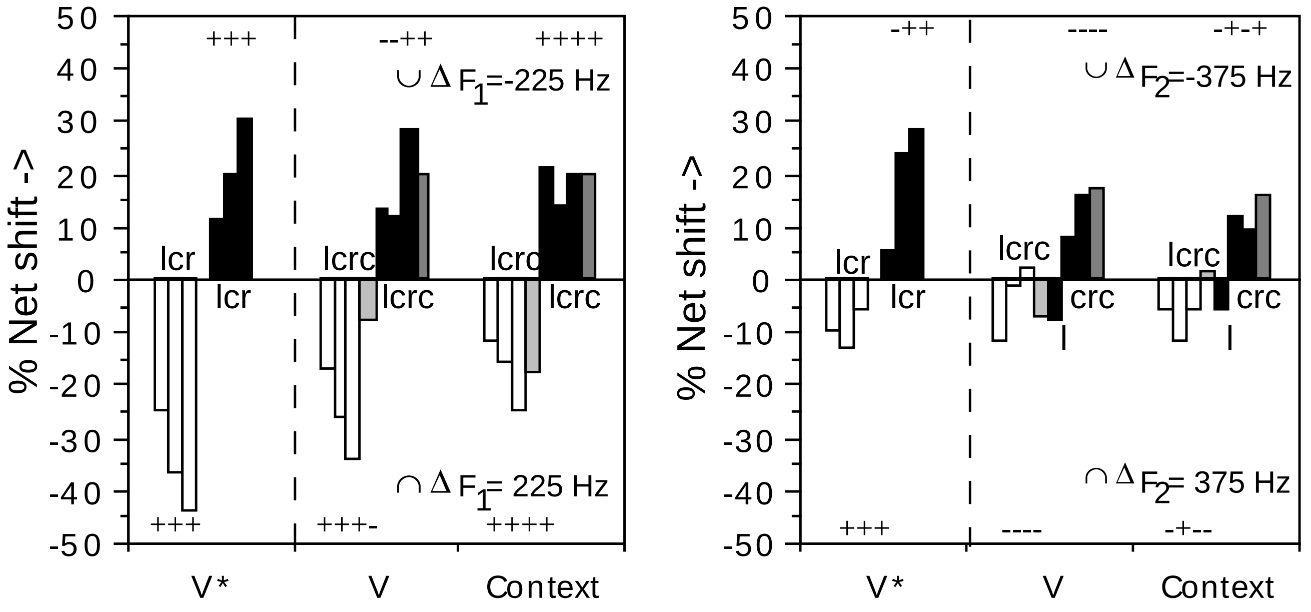
Both theories have problems with some data (proponents of both theories have thoroughly disproved each other's point).

Example: Target overshoot ?









Vowel identification experiments
V*, V: Isolated synthetic vowels (two experiments)
Context: synthetic /nVf/, /fVn/ pseudosyllables
+: $p<0.001$ two tailed sign test

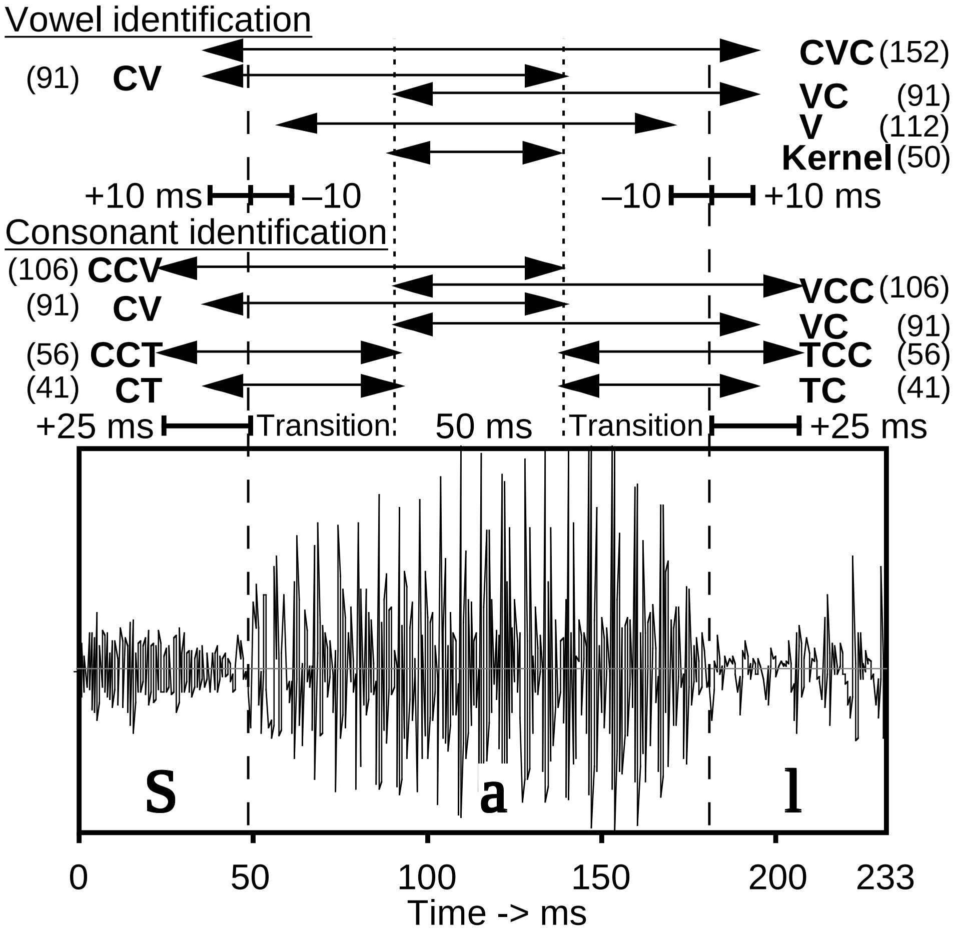
Pattern-recognition models of phoneme recognition

Strong theories (classical theories)
Presuppose strong (fixed) links between the symbolic (phoneme) and acoustic level.
Strong theories of phoneme perception localize acoustic information inside the segment proper. Context information is *always* redundant.

Weak theories
Map cues directly to phoneme sized categories. Allow any regularity to be used for recognition. (Nearey, 1992, 1997)
Weak theories suppose that any speech can contain new (unique) information, even if it originates from the local "context".
Weak theories fit in a Pattern-recognition framework. (Smits, 1997)

Strong theories of phoneme recognition (e.g., motor theory) tend towards an obligatory phoneme hypothesis.
Weak theories of phoneme recognition tend towards a lax phoneme hypothesis

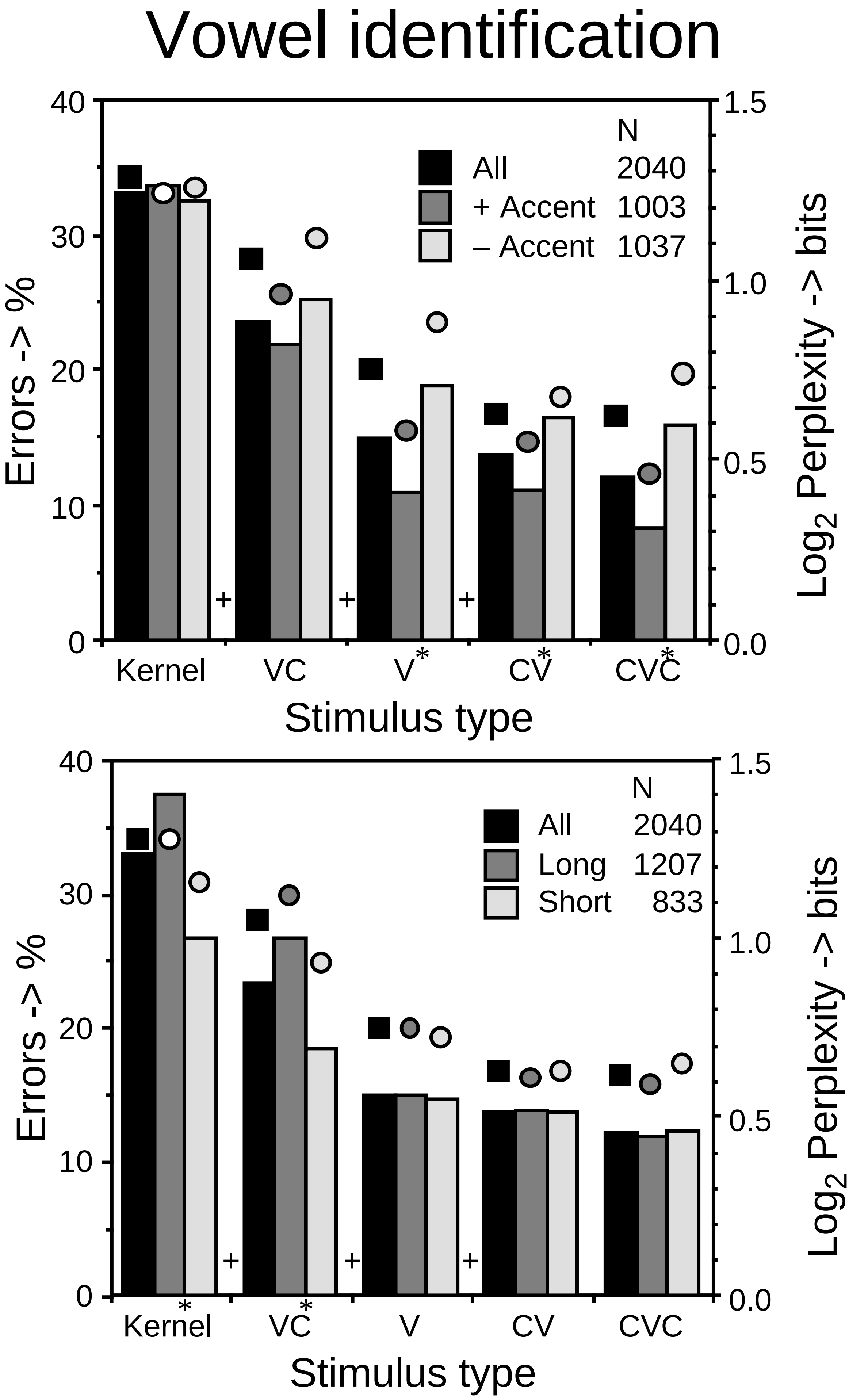
Example of contextual effects on phoneme recognition (gating task)



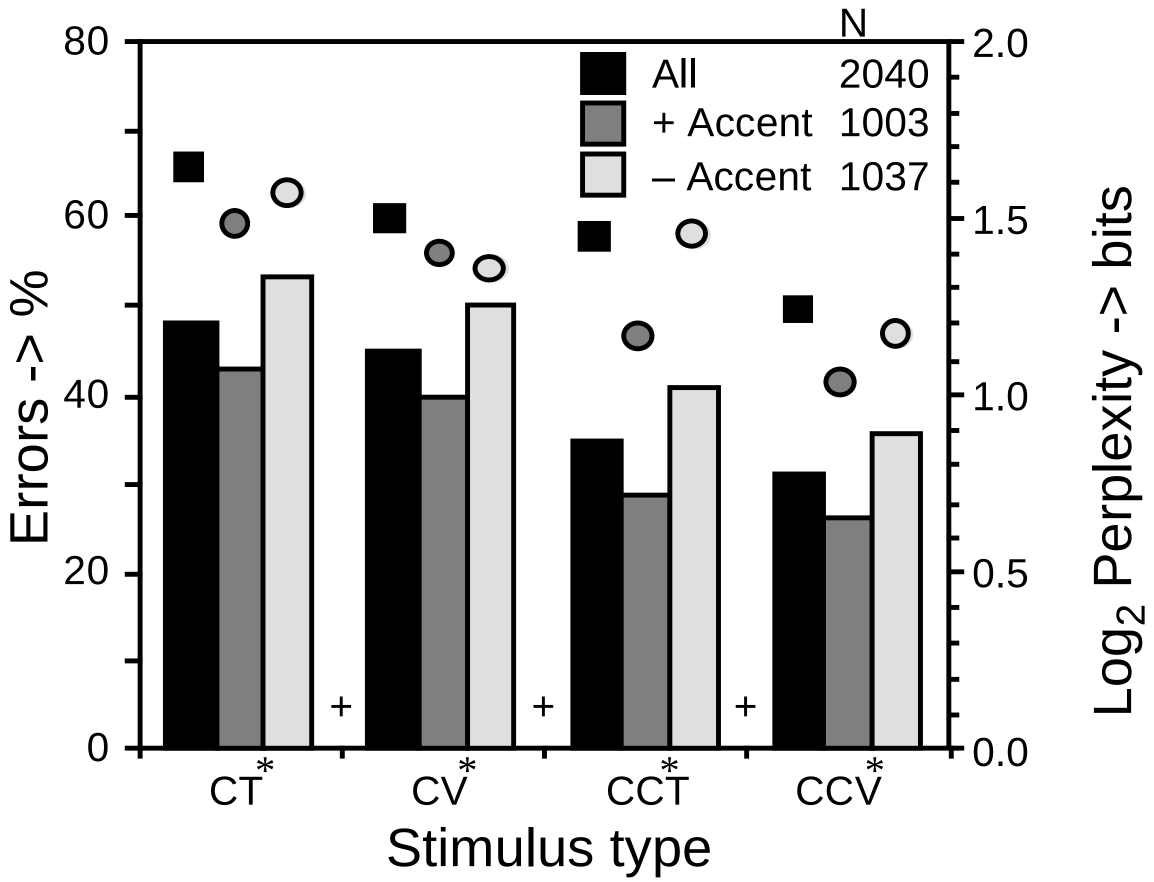
Vowel identification:	
Kernel	50 ms
Kernel+transitions (V),	~ 110 ms
Consonant+transition+Kernel (CV)	~ 90 ms
Kernel+transition+Consonant (VC)	~ 90 ms
Consonant+Vowel+Consonant (CVC)	~ 152 ms

*Pre*vocalic *consonant* identification (C=short/CC=long fragment):	
consonant fragment+transition (CT/CCT)	~ 40/55 ms
consonant fragment+transition+kernel (CV/CCV)	~ 90/105 ms

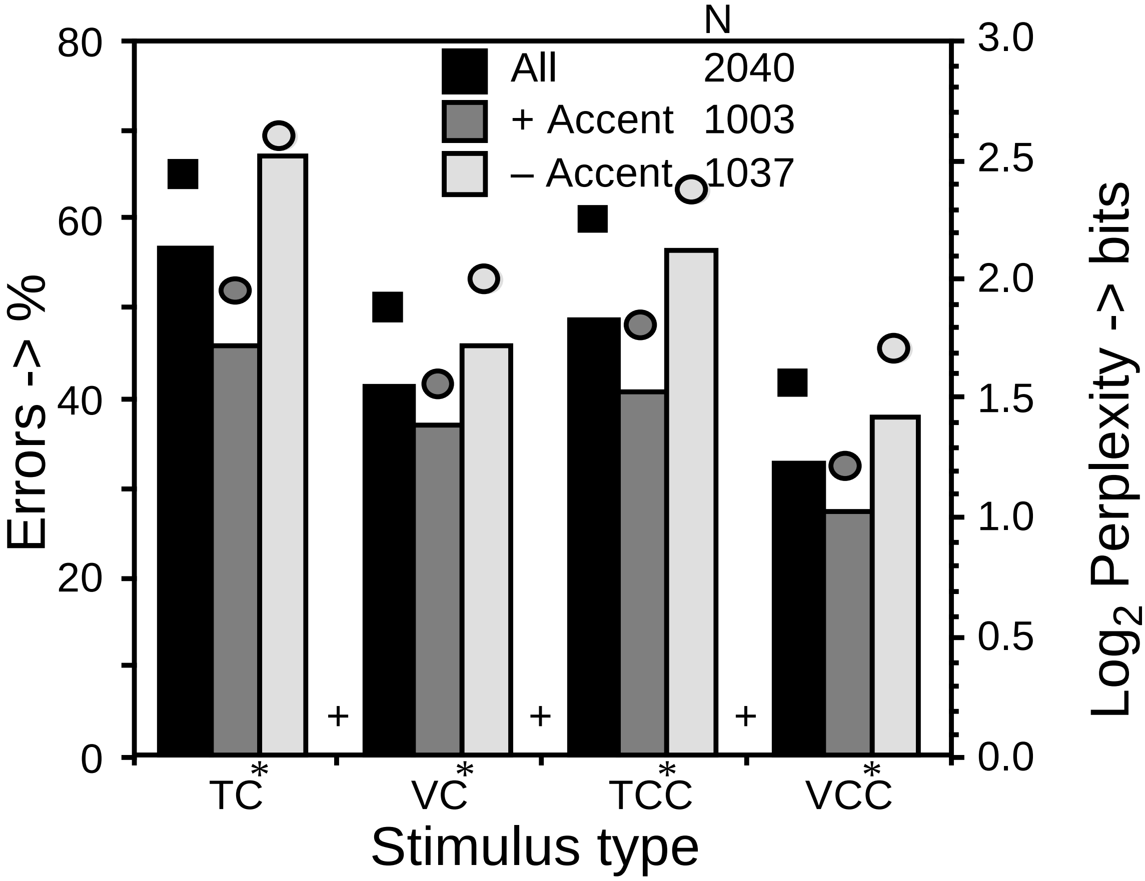
*Post*vocalic *consonant* identification (C=short/CC=long fragment):	
transition + consonant fragment (TC/TCC)	~ 40/55 ms
kernel+transition+consonant fragment (VC/VCC)	~ 90/105 ms


Vowel identification
Errors -> %
Log$_2$ Perplexity -> bits
0
10
20
30
40
0.0
0.5
1.0
1.5
N
All 2040
+ Accent 1003
– Accent 1037
+
+
+
Kernel
VC
V*
CV*
CVC*
Stimulus type
Errors -> %
Log$_2$ Perplexity -> bits
0
10
20
30
40
0.0
0.5
1.0
1.5
N
All 2040
Long 1207
Short 833
+
+
+
Kernel*
VC*
V
CV
CVC
Stimulus type

Pre-vocalic Consonant identification



Post-vocalic Consonant identification



Gating conclusions

1 Phoneme identification benefits from all speech including speech from neighbouring phonemes

2 Speech preceding the target fragment provides more benefits to recognition than speech following it

From **2** we can conclude that phoneme recognition (phoneme naming) is a fast process, the labeling is concluded when the "isolation point" is reached.

Phonemes in context

A reanalysis of 'Bad-Bet' type of experiments pointed out the importance of the perceived *identity* of a neighboring phone/phoneme for recognition. (Nearey 1990)

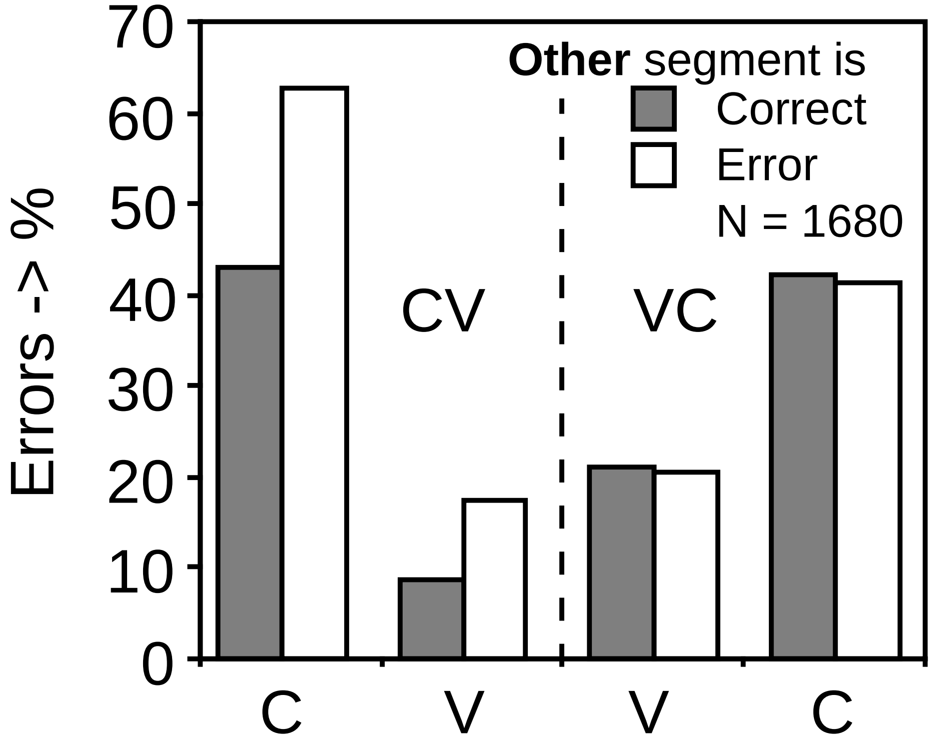
A reanalysis of classical studies has shown that all studies that claimed some kind of "dynamic specification" could not distinguish parameter extrapolation from phonemic context effects.

Only when the appropriate context was heard, did the subjects "compensate" for coarticulation/reduction. The "extent" of compensation was independent of the specific parameter contours.

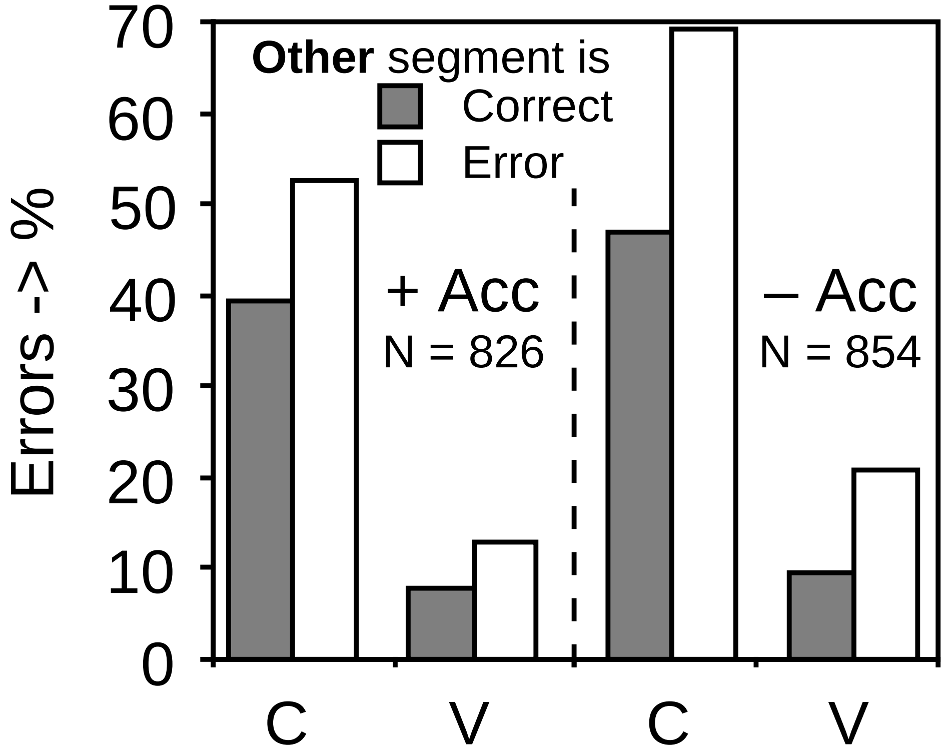
All results (as far as I know) can be explained by a mechanism in which the PHONEMIC context is used to interpret the target PHONE.

Phonemic context

Vowel and consonant recognition
CV versus VC tokens



CV tokens only



Task effects: Parallel processing

Phoneme monitoring

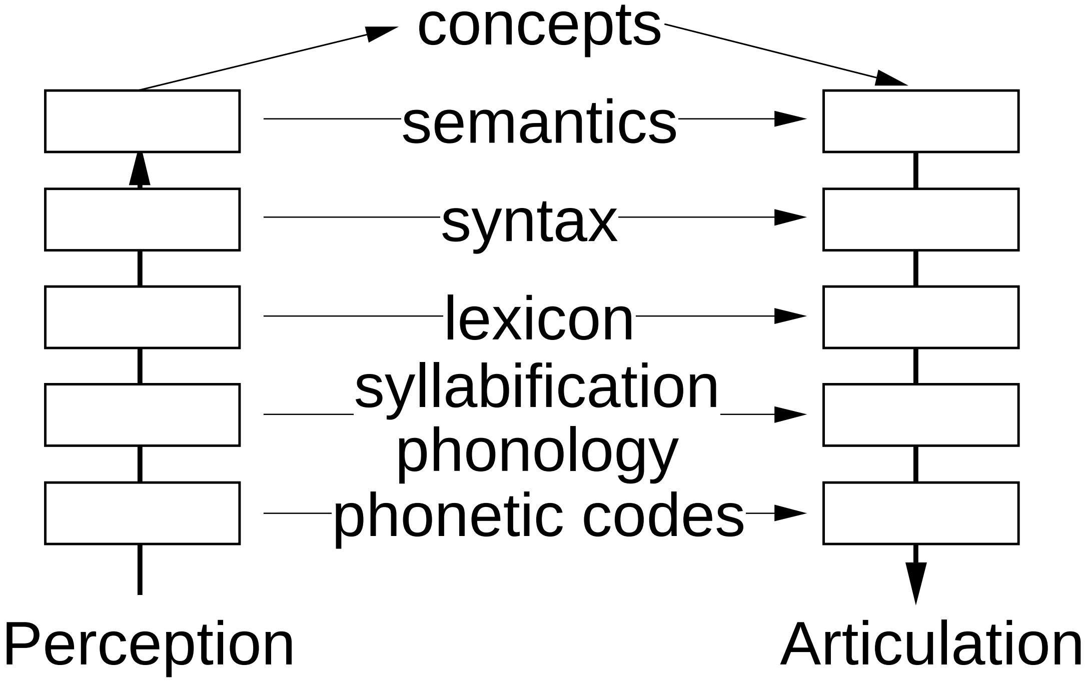
Transitional probability (tri/diphone freq.) affects phoneme (**C**) monitoring in "difficult" CV**C**C , but not in "easy" CV**C**_ tokens.
(McQueen and Pitt, ICSLP 1996, 2502-2505)

Shadowing

Close shadowers react fast (~250 ms delay) before they actually understand the words.

Delays are affected by task variables which change phonological, lexical, syntactic, and semantic "interference".
(Marslen-Wilson, SpeCom 4, 1985, 55-73)

Monosyllabic words from mixed word lists induce larger delays than syllables from pure syllable lists.
(297 ms vs. 258 ms, for delays < 400 ms)



concepts

semantics

syntax

lexicon

syllabification
phonology

phonetic codes

Perception

Articulation

Other Aspects of phone categorization

Initial categorization is non-exclusive:

- Ganong effect
- Phonemic restoration
- Sublabelling in categorical perception (Van Hessen and Schouten, 1992)

Categorization is Bottum Up

- Uses "Bayesian like" rules for integration (Norris, McQueen and Culter,2000)

- McGurk effect (Massaro and Friedman 1990)

A synthesis?

Phoneme recognition is a pure bottum up process. (Norris, McQueen and Cutler, 2000)

Phoneme recognition fits a "weak", pattern matching framework. (Smits, 1997, Nearey, 1992, 1997)

Phoneme recognition is lax.

Phoneme recognition starts with a phone categorization process that :
- recycles cues
- combines all information (Bayesian decissions?)
- preserves ambiguities
  (all possible categories are available)

The result of the categorization can be tought of as a lattice(?) of phone categories that can be fed into the lexicon (word recognition, phoneme identification or monitoring) or the production apparatus (shadowing).

The next stage will reduce the initial lattice to a single representation, according to the task at hand.

Unanswered questions about phone categorization

Is the initial categorization really a distinct process or just an integral part of the lexical or motor route (or both)?

What is the nature of the initial categories, e.g., (allo)phones or phonemes?
Are they real?

Are several phone categories "activated" in parallel (a lattice) or is this an artefact of experimental manipulations?

Is there an "isolation point" for phoneme naming or are label decisions forced by processing or temporal constraints?