# INTER-DEPENDENT CATEGORIZATION OF VOICES AND SEGMENTS

*Anne Cutler*[a], *Attila Andics*[a] *& Zhou Fang*[b]

[a]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands;
[b]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, the Netherlands
Anne.Cutler@mpi.nl/a.cutler@uws.edu.au; attila.andics@gmail.com; y.fang@student.ru.nl

## ABSTRACT

Listeners performed speeded two-alternative choice between two unfamiliar and relatively similar voices or between two phonetically close segments, in VC syllables. For each decision type (segment, voice), the non-target dimension (voice, segment) either was constant, or varied across four alternatives. Responses were always slower when a non-target dimension varied than when it did not, but the effect of phonetic variation on voice identity decision was stronger than that of voice variation on phonetic identity decision. Cues to voice and segment identity in speech are processed inter-dependently, but hard categorization decisions about voices draw on, and are hence sensitive to, segmental information.

**Keywords:** voice, vowels, consonants, Garner task

## 1. INTRODUCTION

The processing of speech delivers information on multiple levels, and listeners appear capable of using any and all levels of information from which they can benefit. One of the types of information presented in speech signals is the talker's voice, which allows listeners to recognize talker identity, or derive relevant speaker-specific information even from the speech of unfamiliar talkers.

Much recent evidence suggests that information about voice is not discarded by normalization processes, but is processed together with phonetic and lexical cues to the spoken message. Indeed, listeners use knowledge of talkers in recognition in many ways [9, 10, 15], suggesting that phonetic and voice processing are highly inter-dependent.

A classic method for testing inter-dependence (or independence) of levels of processing is the selective attention paradigm, or Garner task [7]. Participants make decisions on (usually: categorize) stimuli that vary in one dimension, while variation on a different dimension is either present or absent – for instance, categorizing stimuli as beginning /b/ or /d/, given tokens with varying vowels (*ba, bi, di*) versus a constant vowel (*ba, ba, da*). Variation in irrelevant dimensions tends to slow target dimension decisions.

The paradigm's attraction is that the decision is simple (two-alternative forced choice: 2AFC) and participants' responses soon stabilise at a fast and consistent level. If, in such a situation, participants cannot selectively attend to one dimension and ignore variation in another, then processing of the two dimensions is likely to be inter-dependent in general. In contrast, if selective attention to a dimension is possible, i.e., decisions are unaffected by variation in another dimension, then the two dimensions can clearly be processed independently.

The task was developed for the study of visual perception (decisions about horizontal and vertical position of a dot proved dependent; decisions on size of a circle and angle of a line through it proved independent [8]). It was rapidly deployed in research on speech perception [4, 12, 19, 20]. In these studies the structure of the visual experiments (two levels each of two dimensions) was applied to auditory categories. A central issue at the time was whether auditory processing needed to be completed before phonetic processing could begin. If so, then auditory categorization should be impervious to phonetic variation, but phonetic classification would always be influenced by irrelevant auditory variation.

In fact, most results showed mutual dependence: whatever the dimensions, irrelevant variation slows RTs. Consonant choices in CV syllables are slowed by vowel variation, and vice versa [20]; vowel choice is affected by pitch or loudness variation, and vice versa [12]; choice of place of articulation is slowed by variation in manner of articulation, and vice versa [4]; choice between lexical tones on CV syllables is slowed by either vowel or consonant variation, and vice versa [11, 18]; choice between two nonsense forms is slowed by varying stress [16].

But dependencies were not always symmetrical, indicating that some processes drew on information from other signal dimensions; place decisions were affected by manner variation more than vice versa

[4], and consonant decisions were affected by pitch variation more than vice versa [11]. Hard decisions were most susceptible to irrelevant variation, while easier decisions were less affected [3, 4, 18, 19].

Reflecting the origin of the selective attention method in the psychophysical tradition [7], these experiments, with few exceptions, used synthetic stimuli in which all except the two dimensions of interest was held constant. This has clear advantages for interpreting findings, but is difficult to apply to multidimensional categories or complex dimensions such as timbre. Talker voice is such a category. Fortunately, result patterns in studies using natural materials e. g., [16, 18] do not noticeably differ from those in the studies with synthesized stimuli.

Mullennix and Pisoni [13] compared phonetic to voice processing, with naturally spoken real-word stimuli. This study was more complex than prior studies, involving up to 16 levels per dimension. Participants categorized words (e.g., *bad*, *pad*, *pill*, or *bill*) as beginning /b/ or /p/, or they classified the voice speaking them as of a male or a female talker.

Overall, RTs were faster in the talker decisions than in phonetic decisions. Mutual dependence again appeared, in that variation in either non-target dimension affected judgements on the other dimension. Crucially, the effects were asymmetric: phonetic decisions were affected more strongly by variation in number of talkers than voice decisions by variation in number of words used.

Listeners' voice processing, however, goes far beyond male-female judgement. Knowledge about talker identity informs phonetic and lexical processing [9, 10, 15]; phonetic category boundaries are rapidly adjusted to cope with talker-specific pronunciations [5, 14]. In the present study a classic Garner experimental design is used with a more challenging, but at the same time natural, voice identity categorization task, along with a phonetic categorization task that is also more challenging than syllable-initial stop consonant categorization. Listeners performed 2AFC between (a) two quite similar male voices, and either (b) two post-vocalic alveolar consonants differing in manner, or (c) two preconsonantal central short vowels. The non-target dimension was either constant, or varied.

Although new voices can be easily learned [2], establishing a new category (here, for the two voice identities) will, we predict, always result in longer RTs and higher error rates than recognizing a known category (here, the segments). The principal issue concerns variation interference: does it occur in each direction, and if so, are the effects symmetrical or not? We predict that voice identity categorization will be harder, and hence show greater interference, than segment categorization.

## 2. EXPERIMENT

### 2.1. Participants

30 native Dutch-speaking University of Nijmegen undergraduates took part in return for a small remuneration. (Results of six further participants were lost due to equipment malfunction).

### 2.2. Stimuli

For [1, 2], a quite homogenous set of young male non-smoking native speakers of Dutch without speech problems or recognizable regional accents (age range: 18-30) had recorded multiple tokens of the eight syllables *met* [mɛt], *mes* [mɛs], *mot* [mɔt], *mos* [mɔs], *let* [lɛt], *les* [lɛs], *lot* [lɔt] and *los* [lɔs]. Four speakers with similar F0 range were selected. The recordings, which had been sampled at 44100 Hz, 16 bits per sample, and equalized for average amplitude, were truncated to VC syllables by excising the initial consonant, using PRAAT. For two speakers, who were used as target voices, six tokens of each of the VC syllables *et*, *es*, *ot* and *os* were chosen, and for the remaining two (non-target) speakers, six tokens each of *et*, *es* and *os*. Segment durations were measured for all tokens.

Two sets of four experimental blocks were constructed. One set contained two blocks each of vowel or voice targets, the other contained two each of consonant or voice targets. In each two blocks with the same target type, one block had a constant context (the same syllable, for voice targets, or the same voice, for segment targets), while the other had a varying context. For vowel decisions, the consonant was always /s/, and for consonant decisions, the vowel was always /ɛ/. The constant syllable for voice decisions was *ot*, and the constant voice for segment decisions was "Peter" (see below); the varying context for voice decisions was all four syllables, and the varying context for segment decisions all four voices. There were 80 trials per block: 16 practice and 64 experimental trials (32 with each target alternative).

### 2.3. Procedure

Participants were seated in a sound-attenuated booth with a monitor and a two-key response box in front of them. They heard the stimuli binaurally over Sennheiser headphones. 16 participants heard the set of four vowel/voice blocks, 14 the

consonant/voice blocks. Order of context blocks (constant, varying) was counterbalanced within each group. Before each block, instructions on the monitor stated the 2AFC for the block; once a key-press started the block, the two alternatives appeared on the screen, in red and green, colour-co-ordinated to the response keys. The voices were named "Peter" and "Thomas", and participants were instructed to guess on an initial response and then to try to learn the voices as rapidly as possible. Feedback was given throughout the experiment; for correct responses, the monitor showed the RT; for incorrect responses, the correct category name appeared. Stimulus presentation and data recording were under control of a computer running Presentation software.

## 2.4. Results

The grand mean correct RT was 438 ms, and the grand mean error rate 15.2%. Table 1 shows the means as a function of target type and context.

### *2.4.1. Error rates*

Segment decisions led, as predicted, to fewer errors than voice decisions ($p < .001$ for both vowel and consonant groups). New-category learning was revealed by a decrease in error rate of voice decisions across voice-target blocks (chiefly borne by a decrease from the first quadrant of each block to the remaining three quadrants); there was no change across quadrant for segment decision error rates (interaction of quadrant and target type $p<.02$). However, the error rates showed no effect at all of context (varying versus constant; $F<1$). Note here that many earlier studies also observed no Garner effect in error rates, and that for segment decisions error responses were much faster than correct responses (mean error RT for decisions on vowels 262 ms, on consonants 343 ms), though correct and error voice RTs did not significantly differ.

**Table 1:** Mean correct RT (ms), in bold, and error rate (%) as a function of target type (segment, voice) and context (constant, varying) for groups whose segment targets were respectively vowels and consonants.

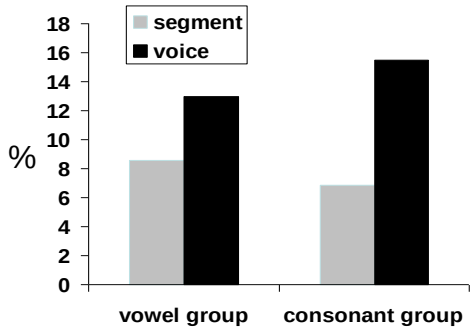
| | Context: | Constant | Varying |
|---|---|---|---|
| Vowel group | Segment targets | **301** 4.5 | **328** 5.1 |
| | Voice targets | **467** 25.1 | **532** 22.7 |
| Consonant group | Segment targets | **407** 5.8 | **428** 5.2 |
| | Voice targets | **481** 26.8 | **562** 26.3 |

### *2.4.2. RTs*

RTs were, also as predicted, faster to segment than to voice targets ($F [1,15] = 79.06$, $p < .001$, $F[1,13] = 30.62$, $p<.001$ for the vowel and consonant groups respectively). Learning effects appeared in the RTs across blocks for segments ($F [3,84] = 3.92$, $p<.02$) and, more strongly, for voices ($F [3,84] = 5.08$, $p<.003$); this was again in each case carried by a decrease from the first to the remaining quadrants). Vowel decisions, as can be seen in the table, were faster than consonant decisions, but this is due to RTs being measured from VC syllable onset; the stimulus measurements revealed an average vowel duration of 111 ms, almost exactly matching the vowel-consonant RT difference (104 ms).

Most importantly, and clearly visible in Table 1, there was a significant "Garner effect", with faster RTs in constant than in varying contexts ($F [1,28] = 61.21$, $p<.001$). This also interacted with the target type comparison ($F [1,28] = 14.7$, $p<.001$). Separate analyses showed a significant effect of context both for segment ($F [1,28] = 22.58$, $p<.001$) and for voice decisions ($F [1,28] = 39.77$, $p<.001$); the interaction, therefore, reflects the fact that the effect is asymmetric. Note that the asymmetry is in the opposite direction to that found by Mullennix and Pisoni [13] for voice decisions amounting to a male-female judgement. Figure 1 shows the effect size per condition; it can be seen that the larger effect of varying segmental context on voice decisions, compared with the effect of voice variation on segment decisions, appears for both groups (the three-way interaction of target type, context and vowel/consonant group did not reach significance).

**Figure 1:** "Garner effect" (RT in varying context minus RT in constant context, as percentage of condition mean), separately for decisions on segment versus voice identity, for the groups with vowels or with consonants as segment targets. Consonant RTs are here normalized for duration of the preceding vowel.



## 3. DISCUSSION

The processing of voice information and segment information is inter-dependent. Irrelevant variation in either of these dimensions complicates simple two-alternative choices in the other. But the size of the effect in the two directions is asymmetric.

This asymmetry and that observed by Mullennix and Pisoni [13] differ. They found voice decisions to be less affected by consonant variation than vice versa, and voice decision to be faster than segment decision; we found in each case the reverse.

The differences reflect the nature of the tasks, however; the results are rather complementary than contradictory. Their voice task (male or female?) was a familiar one, and hence easier than our task of learning which of two quite similar new voices was "Peter" and which "Thomas". Male-female decision can also largely be made on the basis of voice F0; lesser effects of phonetic on F0 decisions than vice versa had emerged earlier [11, 19]. Mullennix and Pisoni's two findings (faster voice than segment RTs; larger Garner effects in segment decisions) are not unrelated, since, as earlier studies also showed, Garner effect size is greater the harder the decision type [3, 4, 18]. This decision difficulty effect is non-trivial; the longer a decision takes, the more scope there is for context effects to be exercised on that decision.

We chose to make all choices quite difficult. The consonants /s/, /t/ exert relatively little coarticulatory effect in a prior vowel, so that early decisions are unlikely; the vowels /ɛ/, /ɒ/ in Dutch contrast only in backness, not in height or duration. But still, the segment contrasts are known, while the voice contrasts were unfamiliar. As the error rates and RTs proved, voice decision was indeed harder.

Hard or not, though, voice categorization is done by listeners every day. Voices can be identified from among many known talkers, and new voices can be learned, if necessary without added information – e.g., we generally track a radio discussion between previously unknown talkers without difficulty. Voice learning is rapid and robust [1], and can be achieved on minimal information [6]. Our results suggest that phonetic information helps identify newly learned voices, thus adding to previous demonstrations [2, 17] of a dependence of voice processing on phonetic information. At the same time we cannot ignore the evidence that phonetic processing also draws on voice information [9, 10, 15]. The two information types are fully interwoven in the speech signal, and are processed in a fully inter-dependent manner.

## 4. ACKNOWLEDGEMENTS

AC is also affiliated to MARCS Auditory Labs, Univ. of Western Sydney, and AA to MR Research Center, Semmelweis Univ., Budapest.

## 5. REFERENCES

[1] Andics, A., McQueen, J.M. Submitted. *Plasticity, robustness and abstraction in voice identity learning.*

[2] Andics, A., McQueen, J.M., van Turennout, M. 2007. Phonetic context influences voice discriminability. *Proc. 16th ICPhS* Saarbrücken, 1829-1832.

[3] Carrell, T.D., Smith, L.B., Pisoni, D.B. 1981. Some perceptual dependencies in speeded classification of vowel color and pitch. *Perc. Psychophys.* 29, 1-10.

[4] Eimas, P.D., Tartter, V.C., Miller, J.L., Keuthen, N.J. 1978. Asymmetric dependencies in processing phonetic features. *Perc. Psychophys.* 23, 12-20.

[5] Eisner, F., McQueen, J.M. 2005. The specificity of perceptual learning in speech processing. *Perc. Psychophys.* 67, 224-238.

[6] Fellowes, J.M., Remez, R.E., Rubin, P.E. 1996. Perceiving the sex and identity of a talker without natural vocal timbre. *Perc. Psychophys.* 59, 839-849.

[7] Garner, W.R. 1974. *The Processing of Information and Structure*. Potomac Md: Erlbaum.

[8] Garner, W.R., Felfoldy, G.L. 1970. Integrality of stimulus dimensions in various types of information processing. *Cog. Psychol.* 1, 225-241.

[9] Hay, J., Warren, P., Drager, K. 2006. Factors influencing speech perception in the context of a merger-in-progress. *J. Phonetics* 34, 458-484.

[10] Johnson, K. 1990. The role of perceived speaker identity in F0 normalization of vowels. *JASA* 88, 642-654.

[11] Lee, L., Nusbaum, H.C. 1993. Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perc. Psychophys.* 53, 157-165.

[12] Miller, J.L. 1978. Interactions in processing segmental and suprasegmental features of speech. *Perc. Psychophys.* 24, 175-180.

[13] Mullennix, J.W., Pisoni, D.B. 1990. Stimulus variability and processing dependencies in speech perception. *Perc. Psychophys.* 47, 379-390.

[14] Norris, D.G., McQueen, J.M., Cutler, A. 2003. Perceptual learning in speech. *Cog. Psychol.* 47, 204-238.

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

[16] Pallier, C., Cutler, A., Sebastián-Gallés, N. 1997. Prosodic structure and phonetic processing: A cross-linguistic study. *Proc. Eurospeech97* Rhodes, 2131-2134.

[17] Remez, R.E., Fellowes, J.M. 1997. Talker identification based on phonetic information. *J. Exp. Psychol.: Hum. Perc. Perf.* 23, 651-666.

[18] Repp, B.H., Lin, H.-B. 1990. Integration of segmental and tonal information in speech perception: A cross-linguistic study. *J. Phonetics* 18, 481-495.

[19] Wood, C.C. 1974. Parallel processing of auditory and phonetic information in speech discrimination. *Perc. Psychophys.* 15, 501-508.

[20] Wood, C.C., Day, R.S. 1975. Failure of selective attention to phonetic segments in consonant-vowel syllables. *Perc. Psychophys.* 17, 346-350.