# Asymmetries in the exploitation of phonetic features for word recognition

**Alexander Martin and Sharon Peperkamp**
*Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Études Cognitives, École Normale Supérieure – PSL Research University, 29 rue d'Ulm, 75005 Paris, France*
*alexander.martin@ens.fr, sharon.peperkamp@ens.fr*

**Abstract:** French listeners' reliance on voicing, manner, and place was tested in a mispronunciation detection task. Mispronounced words were more likely to be recognized when the mispronunciation concerned voicing rather than manner or place. This indicates that listeners rely less on the former than on the latter for the purposes of word recognition. Further, the role of visual cues to phonetic features was explored by the task being conducted in both an audio-only and an audiovisual version, but no effect of modality was found. Discussion focuses on crosslinguistic comparisons and lexical factors that might influence the weight of individual features.


[AC]
**Date Received:** December 10, 2014 **Date Accepted:** March 17, 2015

## 1. Introduction

The phonetic features that speech sounds are composed of have long been known to play a role in both speech production and speech perception. In speech production for instance, speech errors can target individual features (Fromkin, 1971), and featurally similar sounds are more likely to interact in speech errors than featurally dissimilar sounds (Stemberger, 1991). Concerning speech perception, identification errors in noise tend to preserve features of misheard sounds (Miller and Nicely, 1955), and consonant identification in dichotic listening is impaired when the two consonants are featurally dissimilar compared to when they are similar (Studdert-Kennedy *et al.*, 1972).

In the realm of word recognition, previous research has investigated the limits of listeners' capacity to recognize words with deformed featural information, using stimuli with deliberate mispronunciations and a variety of tasks. This research has shown first and foremost that the number of feature changes between a word's correct pronunciation and a mispronounced variant is crucial in its recognizability. For instance, written words are primed by auditorily presented mispronunciations of semantically related words, but only if the mispronunciations differ in at most two features (Connine *et al.*, 1993). This is hardly surprising, as the more a word's pronunciation is altered, the more difficult it should become to recognize. But what about the differential perceptual weight of the features themselves?

Cole *et al.* (1978) argue that voicing is indeed less important compared to place. They asked participants to detect mispronunciations of words in 20-min-long stories. Words could be mispronounced in one or more features. In one experiment, mispronunciations concerned either the voicing or the place of the initial consonant. Changes in place elicited higher detection rates than changes in voicing. This effect was robust across the different consonants, indicating that the features were processed similarly regardless of the consonant to which they were associated. Note, though, that as participants could rely on the sentential context, the results might partly reflect the predictability of the particular words used in the experiment.

More recently, Ernestus and Mak (2004) used mispronounced words in isolation, avoiding biases introduced by sentential context. In their study, Dutch listeners

performed a lexical decision task in which they had to reject words whose initial stop or fricative was mispronounced in voicing, place, or manner. No differences in error rates according to the type of feature change were observed for mispronunciations of stop-initial words. For fricative-initial words, however, error rates were higher for voicing mispronunciations than for place or manner mispronunciations. Ernestus and Mak (2004) argued that these results reflect the fact that in Dutch, the voicing feature is relatively uninformative in word-initial fricatives. Indeed, fricatives (but not stops) are subject to phonological processes that change voicing word-initially, and in some varieties of Dutch, all word-initial fricatives are realized as voiceless. Basing themselves on their knowledge of the phonological processes of their language, listeners would therefore pay less attention to voicing than to place and manner, but only in fricatives.

It should be noted that reasoning by Ernestus and Mak (2004) fails to explain the data from Cole *et al.* (1978), who observed an asymmetry between place and voicing in English listeners. Indeed, neither voicing nor place is affected word-initially by any phonological process in English. The aim of the present study is therefore to revisit the issue of relative weight of phonetic features in word recognition. Our case study concerns word-initial obstruents in French. The French obstruent inventory provides a particularly good test case to explore the various weights of phonetic features, as its members are divided evenly over two manners of articulation, three places of articulation, and two voicing values (Table 1). Thus, a change in any of the three features of an obstruent yields another obstruent.

Contrary to Dutch, French has no phonological processes affecting the class of word-initial obstruents or any of its subsets. Thus, if listeners weight their attention to individual features as a function of their informativity in the sense of Ernestus and Mak (2004), we should observe no differences in French listeners among voicing, place, and manner mispronunciations.

Furthermore, we add a novel component by exploring the role that visual cues may play in word recognition. This aspect is particularly relevant, insofar as the place feature has salient cues in the visual signal. We expect that, compared to audio-only input, audiovisual input should make mispronunciations of place—but not of manner or voicing—more disruptive for word recognition. We therefore conducted our experiment with two sets of stimuli: an audio-only set and an audiovisual one.

We report on an experiment in which participants either heard or both saw and heard a speaker produce a series of real, correctly pronounced words as well as mispronounced words interspersed among clear non-words. Their task was to detect both correctly pronounced and mispronounced words; the latter differed from their correctly produced counterparts in one feature of their initial phoneme, which was always an obstruent.

## 2. Methods

### *2.1 Stimuli*

For all of the French obstruents but /z/, we selected three disyllabic French words which (1) contained no other obstruent, (2) yielded a non-word if any one of the

Table 1. The 12 French obstruents arranged vertically by place (in bold) and horizontally by manner (italics) and voicing.

| | *Plosive* | | *Fricative* | |
|---|---|---|---|---|
| | Voiceless | Voiced | Voiceless | Voiced |
| **Labial** | p | b | f | v |
| **Coronal** | t | d | s | z |
| **Dorsal** | k | g | ʃ | ʒ |

features (voicing, manner, or place) was modified in the initial obstruent (e.g., the real word /deli/ "misdemeanor," is turned into the non-words /teli/ through a voicing change, /zeli/ through a manner change, and /beli/ and /geli/, through place changes), and (3) had a higher frequency than all of their phonological neighbors, according to the Lexique 3.80 database (New *et al.,* 2001). For /z/, only two such words were selected, as the French lexicon does not have a third one satisfying the selection criteria. Frequency of the words ranged from 0.06 to 186 occurrences per million tokens (mean: 24).

For each of these 35 base items, we created mispronunciations that were non-words by changing one feature of the initial obstruent. Each base item thus yielded four mispronunciations, one with a voicing change, one with a manner change, and two with a place change. An additional 127 non-word fillers were randomly generated which were also disyllabic, contained only one, initial, obstruent, and had no real word phonological neighbors.

All base items, mispronunciations, and fillers were recorded individually by a female native speaker of French in a soundproof booth with an M-Audio Micro Track II digital recorder and an M-Audio DMP3 pre-amplifier in 16-bit mono at a sampling rate of 44.1 kHz. Video of the speaker including her whole face and stopping at her shoulders was simultaneously recorded at 60 frames per second in 720p HD resolution. The speaker was positioned in the center of the frame with two spotlights cross-positioned to eliminate shadows on her face. The average audio stimulus lasted 517 ms. Video was recorded both before and after the production of the word by our speaker such that each video lasted exactly 1500 ms.

### *2.2 Procedure*

Two versions of the experiment were prepared, one with audio-only stimuli and the other with audiovisual stimuli. Half of the participants were tested with the audio-only stimuli, the other half with the audiovisual stimuli. During the experiment, participants sat in front of a computer screen in a sound-attenuated room while stimuli were played binaurally through a headset. In the audio-only version, participants were presented with a black screen for the entire duration of the experiment. In the audiovisual version, videos were played in synchrony with the audio, depicting the woman producing the word being presented through the headset; in between stimuli, a gray box was displayed over the area of the screen where the speaker was portrayed. Participants were told that a list of items would be read to them by a stroke patient. The patient was said to have reading difficulties and, more specifically, to produce mostly unintelligible words when reading aloud individual nouns, while occasionally producing intelligible words or very close mispronunciations. Participants were asked to press a key (in the audio-only version) or a button (in the audiovisual version) whenever they recognized a noun—whether it was pronounced correctly or incorrectly—(go response) and to do nothing otherwise (no-go response). If a go response was recorded, a dialogue box prompted the participant to report the recognized word by typing it on a computer keyboard. The next stimulus was then played after 1000 ms.

Participants were presented with target stimuli [i.e., the 35 base items presented in the control condition (correct pronunciation) or in one of the three test conditions (voicing, manner, or place mispronunciation)], interspersed among filler stimuli (i.e., clear non-words). Six lists of stimuli were prepared, with each base item appearing only once in each condition and no specific sound manipulation appearing twice. Because of the missing /z/ item, though, not all manipulations were present in each list. Participants were randomly assigned to one of the six counterbalanced lists. As a consequence, subjects heard on average 29 target stimuli (the number varied across participants because of the absence of a third /z/-initial base word), and all 127 filler stimuli. Stimuli were presented semi-randomly, such that target stimuli never directly followed one another, with an ISI of 2500 ms.

The experiment started with a short training phase, containing stimuli recorded by a different speaker than was used in the main task. In this phase, three target stimuli (one correctly pronounced noun and two mispronunciations) were mixed into a dozen filler stimuli. The mispronunciations concerned sonorants rather than obstruents so as not to interfere with the main task. Participants received feedback about their responses; if they failed to identify one of the mispronunciations, a message alerted them to their error and indicated the noun they were meant to identify. They had to correctly identify two out of the three target stimuli before moving on to the main task. If necessary, the training phase was repeated until this criterion was met.

### *2.3 Participants*

Forty-eight native speakers of French, 12 men and 36 women aged between 18 and 32 (mean: 22.5), participated. They were randomly assigned to either the audio or the audiovisual version of the experiment. None of the participants reported any history of hearing problems and they all had corrected-to-normal vision.

## 3. Results

Six participants (two in the audio-only version and four in the audiovisual version) made more than 30% errors on control stimuli (correctly produced real words); their data were excluded from the analyses. Data from the 42 remaining participants were analyzed using generalized mixed models in R (Bates *et al.,* 2014) with a declared binomial distribution given that the dependent variable was binary (hit or miss).

We analyzed individual hit rates on both test items (mispronunciations) and control items (correct pronunciations). For the former, a hit was defined as a go response with reporting of the correct target word (e.g., identification of the mispronunciation /teli/ as target /deli/ "misdemeanor"). The mean hit rates per condition are shown in Fig. 1.

An original model was created with modality (audio-only or audiovisual), condition (correct, voicing mispronunciation, manner mispronunciation, or place mispronunciation), frequency of base item, as well as all interactions as fixed factors, and subject and base item as random factors. Note that frequency was included to control for the large amount of variability in the base words. By-subject random slopes for condition and frequency were included given that each participant was exposed to all four conditions and all frequency values. For the random factor base item, however, only a random slope for condition was included, as frequency perfectly correlates with base item. As neither modality nor frequency was a significant predictor in this model (both $z < 1$), and since they did not interact with any other factors, both were excluded as factors from subsequent models.

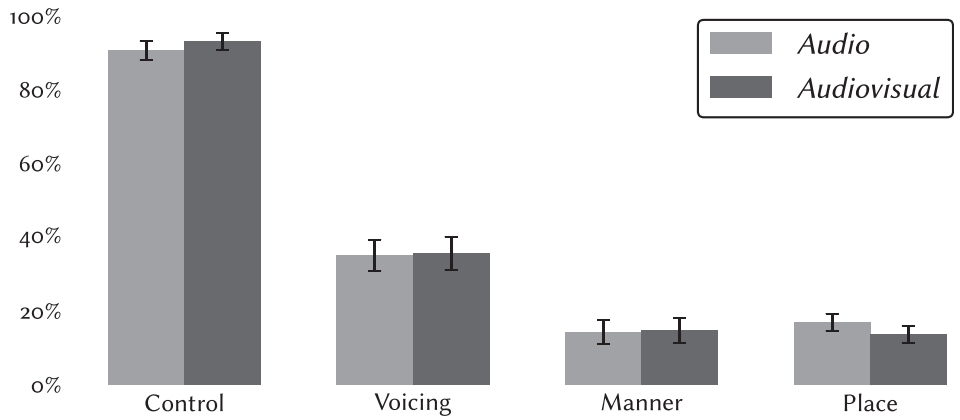


Fig. 1. Hit rates by condition and modality. Error bars represent standard error of the mean.

Our final model included condition as a fixed factor and subject and base item as random factors, each with a random slope for condition. All three mispronunciation conditions were significantly different from the control condition (i.e., correct pronunciation): voicing [$\beta = -4.44$, standard error (SE) $= 0.72$, $z = -6.12$, $p < 0.0001$], manner ($\beta = -6.07$, $SE = 0.79$, $z = -7.69$, $p < 0.0001$), place ($\beta = -6.32$, $SE = 0.76$, $z = -8.33$, $p < 0.0001$). Furthermore, the voicing mispronunciation condition differed significantly from both the place ($\beta = -1.92$, $SE = 0.40$, $z = -4.76$, $p < 0.0001$) and manner ($\beta = -1.65$, $SE = 0.42$, $z = -3.92$, $p < 0.0001$) mispronunciation conditions. However, there was no such difference between the place and manner mispronunciation conditions ($z < 1$).

## 4. Discussion

Using a novel mispronunciation detection paradigm, we examined the relative importance of phonetic features in word recognition in both the auditory and the audiovisual modality. Overall performance in the mispronunciation conditions was low, indicating that participants had a hard time recognizing words with even a one-feature change. Nonetheless, clear differences were found between the types of feature changes, indicating that different features have different degrees of importance for word recognition. Specifically, participants were able to identify words that were mispronounced with a voicing change significantly more often than ones that were mispronounced with a manner or a place change. This held for both the auditory and the audiovisual versions of the experiment. Thus, voicing mispronunciations are less detrimental to word recognition, indicating that listeners give less weight to the voicing feature compared to the place and manner features, regardless of modality.

The absence of an effect of modality was unexpected. To our knowledge, no previous studies have explored multimodal interaction in the processing of mispronunciations, but given that visual information contains cues to place but only secondarily to manner[1] and not at all to voicing, we indeed expected that in the audiovisual version, participants' ability to recognize words with a mispronunciation of the place feature would be even more reduced. It is possible we did not obtain such a difference because of a floor effect; indeed, even in the audio-only version, performance in the place condition was low. Alternatively, it has been argued that visual cues are especially relied upon under difficult listening conditions, for instance when the auditory signal is presented in noise (e.g., Sumby and Pollack, 1954). Yet another potential explanation comes from recent research suggesting that visual input might be involved in prelexical but not in lexical processing (Samuel and Lieblich, 2014). More research is necessary to tease apart these possibilities and to explore the role of visual input in the processing of mispronunciations.

Our finding that voicing is less important than place and manner is in line with Cole *et al.* (1978) but not with Ernestus and Mak (2004). Recall that Cole *et al.* (1978) explored mispronunciations in context, using English. In their study, participants were presented with stories in which they had to detect mispronounced words. Performance was better on words with a mispronunciation of the place than of the voicing feature, indicating that the former hamper lexical access more than the latter. The predictability of the particular words manipulated in the stories might have biased the results, though. In the present experiment, we used isolated words and, moreover, ruled out item-specific effects, as each item was used in all mispronunciation conditions and had a higher frequency than all of its phonological neighbors. This, then, attests to the robustness of the larger weight of both place and manner compared to voicing.

As to Ernestus and Mak (2004), using a lexical decision task with Dutch listeners, they observed no difference between place and manner mispronunciations, and higher error rates for voicing mispronunciations in fricatives but not in stops. The finding that word-initially, voicing is less important than place and manner, but only in fricatives, is accounted for by their hypothesis that a feature's relative weight is determined by its informativity within the language. Indeed, in Dutch, voicing is relatively

uninformative in word-initial fricatives because of a phonological devoicing process. In French, however, all obstruent features are equally informative word-initially, at least according to the definition of Ernestus and Mak, as none of the French obstruent features is modified by a phonological process in word-initial position.[2] Thus, their hypothesis fails to explain the relative importance of place and manner compared to voicing in French.

This does not imply that the idea of a feature's informativity determining its weight should be abandoned altogether. Other factors might influence the informativity and hence the weight of individual features. In particular, a feature's functional load (i.e., the extent to which it is used to distinguish words from one another in the lexicon), might play a role: The higher a feature's functional load, the more we would expect listeners to pay attention to it during word recognition.

Calculations of functional load have traditionally focused on segment pairs rather than individual features (Hockett, 1967), and although more recent formalisms have been adapted to explore featural comparisons (Surendran and Niyogi, 2003), different measures of functional load are still being debated (cf. Wedel *et al.,* 2013). This current state of affairs makes testing the hypothesis difficult. As a first step, however, it would be interesting to compare the relative importance of phonetic features in prelexical versus lexical processing, since lexical factors such as functional load should not affect prelexical perception. Of course, confusion studies such as the one by Miller and Nicely (1955) have already examined the relative importance of features in prelexical perception and have shown that contrary to all findings with lexical tasks, place is more likely to be misperceived than voicing. However, these studies used stimuli presented in noise. As noise masks the spectral properties of sounds differentially, this finding is uninformative for the present research question. Future studies addressing this question should therefore use a prelexical task with stimuli presented in clear speech to determine, all things being equal, which features are more perceptually salient. Such research could reveal differences in featural asymmetries between prelexical and lexical perception and could examine whether such differences stem from influences from the lexicon.

**Acknowledgments**

This work was supported by ANR-13-APPR-0012 LangLearn, ANR-10-LABX-0087, IEC and ANR-10-IDEX-0001-02 PSL. We thank Auréliane Pajani for recording the stimuli, Michel Dutat for technical assistance, and Isabelle Dautriche for help with the statistical analyses.

**References and links**

[1]In particular, labial stops and fricatives visibly differ in that the former are bilabial and the latter labiodental. We do not have enough trials to analyze an effect of modality on labial obstruents only.

[2]Word-finally, the voicing feature is subject to assimilation, but based on their data from Dutch listeners, Ernestus and Mak (2004) argue that a feature's informativity is specific to its position within the word.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2014**). "lme4: Linear mixed-effects models using Eigen and S4," R package version 1.1-7, http://CRAN.R-project.org/package=lme4.

Cole, R. A., Jakimik, J., and Cooper, W. E. (**1978**). "Perceptibility of phonetic features in fluent speech," J. Acoust. Soc. Am. **64**(1), 44–56.

Connine, C., Blasko, D., and Titone, D. (**1993**). "Do the beginnings of spoken words have a special status in auditory word recognition?," J. Mem. Language **32**, 193–210.

Ernestus, M., and Mak, W. M. (**2004**). "Distinctive phonological features differ in relevance for both spoken and written word recognition," Brain Language **90**(1–3), 378–392.

Fromkin, V. (**1971**). "The non-anomalous nature of anomalous utterances," Language **47**(1), 27–52.

Hockett, C. (**1967**). *The Quantification of Functional Load: A Linguistic Problem*, U.S. Air Force Memorandum RM-5168-PR.

Miller, G., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**(2), 338–352.

New, B., Pallier, C., Ferrand, L., and Matos, R. (**2001**). "Une base de données lexicales du français contemporain sur internet: LEXIQUE" ("A lexical database from the French Contemporary online: GLOSSARY"), Ann. Psychol. **101**, 447–462.
Samuel, A. G., and Lieblich, J. (**2014**). "Visual speech acts differently than lexical context in supporting speech perception," J. Exp. Psychol. Hum. Percept. Perform. **40**(4), 1479–1490.
Stemberger, J. P. (**1991**). "Apparent anti-frequency effects in language production: The addition bias and phonological underspecification," J. Mem. Language **30**, 161–185.
Studdert-Kennedy, M., Shankweiler, D., and Pisoni, D. B. (**1972**). "Auditory and phonetic processes in speech perception: Evidence from a dichotic study," Cognit. Psychol. **3**(3), 455–466.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Surendran, D., and Niyogi, P. (**2003**). *Measuring the Usefulness (Functional Load) of Phonological Contrasts*, Department of Computer Science at the University of Chicago, Technical Report No. TR-2003.
Wedel, A., Kaplan, A., and Jackson, S. (**2013**). "High functional load inhibits phonological contrast loss: A corpus study," Cognition **128**(2), 179–186.