Holistic perception of phonological variants

Holistyczna percepcja wariantów fonologicznych

Linda Shockey*, Zinny Bond**

*University of Reading
l.shockey@reading.ac.uk
**Ohio University
bond@ohio.edu

ABSTRACT
The goal of this paper is to show that knowledge of phonology, including phonological reductions, is part of native speaker competence and that to apply this knowledge, it is necessary to assume a perceptual window larger, in some cases much larger, than the segment or syllable. In the second part of the paper, we also examine the ability of non-native-speakers to use the larger patterns in speech perception and ask to what extent these shortcuts can be found in other languages.

STRESZCZENIE
Celem pracy jest pokazanie, że znajomość fonologii, włączając w to redukcje fonologiczne, jest częścią kompetencji mówców natywnych, a dla zastosowania tej wiedzy konieczne jest przyjęcie większego okna percepcyjnego, w niektórych przypadkach znacznie wykraczającego ponad segment czy sylabę. W drugiej części artykułu badamy również zdolność mówców nienatywnych do wykorzystywania w percepcji mowy większych wzorców i stawiamy pytanie, do jakiego stopnia tego rodzaju skróty można znaleźć w innych językach.

1. Sound perception as pattern-recognition

We do not advocate any particular model of speech perception. In common with many current models, we assume that the human brain is quintessentially a pattern-recognition device and that spoken language is composed of a large number of simultaneous and overlapping patterns which are all recognised and used selectively by the brain. Where human beings are superior to computers is in the ability to use pattern matching with reference to a given situation. In speech perception, it is obvious that attention to patterns of all sizes is necessary: small patterns provide the traditional acoustic cues for distinguishing 'peach' and 'teach' and other words with minimal differences ... in other words, a phone-based approach is necessary in some cases. But our research suggests that one or more larger patterns or templates is being used in addition to these, and we suggest that phonological patterns are among them. Correct perception is a result of getting the right perspective on a signal both in the smaller and larger patterns. Once we have matched the right pattern or patterns, we can interpret the 'soundscape'. Similar ideas have been discussed under the heading 'normalisation', and much has been said about the various types. This may thus be considered a new angle on an old question.

Here we are focusing on *phonological* equivalencing, i.e. there are multiple realisations associated with most words and phrases in English, related to each other by learnable sound patterns: the word ‘mountain’ can be pronounced [mã̃ʊ̃ntə̃n, mã̃ʊ̃ntn̩, mã̃ʊ̃nʔn̩, mã̃ʊ̃ʔn̩] or even [mãʊʔ͡pm̩], and the knowledge which allows the mapping into a single item is used effortlessly and unconsciously by most native speakers of English. These variants are outlined in detail in many publications, including Shockey [1]. What is interesting is that they go unnoticed by both the speaker and the hearer, and in fact speakers often hotly deny that they use other than the citation form. It is as if they all contain the same information and are perceptually indistinguishable. A dog with three legs is perceived as having something missing, but the phrase ‘first one’ is not perceived as incomplete when it is pronounced [ˈfɝswʌn]. The profile of the utterance is undamaged, much as when we identify the shapes in Figure 1 as a triangle or as two rectangles. The missing parts don’t count.

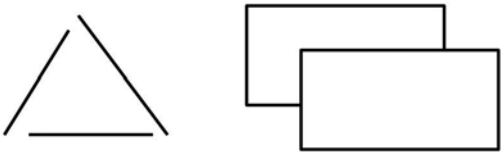

Figure 1: Perceiving incomplete shapes.

We can relate this to the well-known Gestalt principles of grouping, conceived in the early part of the last century to describe the perceptual organization of visual stimuli. Gestalt theorists followed the basic principle that the whole is greater than the sum of its parts. In other words, the whole (a picture, a car) carried a different and altogether greater meaning than its individual components (paint, canvas, brush; or tire, paint, metal, respectively). In viewing the ‘whole’, a cognitive process takes place – the mind makes a leap from comprehending the parts to realizing the whole. In other words, perception is holistic. Gestalt theory is now regarded as overly simple and not adequately explanatory by researchers in the field of vision, and we suggest only that in the case of perception of phonological variants the parallel is striking. It appears that similar larger patterns are being used perceptually. We have evidence from two areas of research: slips of the ear and gating.

2. Part 1: Evidence from Slips of the Ear

Here we argue that the mis-application of a template or pattern shows that it exists, even though it might not be helpful in every case. This happens in other perception modes: as we have seen above, the template for face recognition is very strong. Humans can recognise faces in extremely low-resolution images. A result is that faces are recognised when they arguably shouldn’t be ... as when a decade-old toasted cheese sandwich said to bear an image of the Virgin Mary sold on eBay for $28,000.

A more historical example can be seen in the painting in from the 16th century Italian artist Giuseppe Arcimboldo (Figure 2). Vegetables no longer figure significantly in the second picture, because it stimulates a holistic face-recognition template.

Similarly, perceptual templates for phonological patterns can be inferred when they are misapplied.

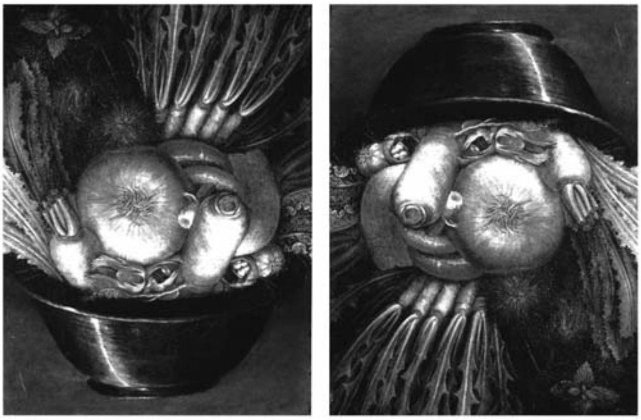

Figure 2: Giuseppe Arcimboldo, *Ortaggi in una ciotola o L'ortolano (Vegetables in a bowl or The Gardener)* (ca. 1580), Museo Civico Ala Ponzone, Cremona, Italy.

2.1. An Elaborated Example

Experienced speakers of casual English show alternations between 'nd' and the nasal by itself, especially word-finally. Listeners therefore have a phonological template which allows the mapping of spoken forms such as 'han (or ham) bone' into 'hand bone'. Clearly this is an operation which needs a suprasegmental window, especially in the case of 'ham', where context allows the decision that the 'm' is really a realization of an 'n'.

Examples of this can be found in slips of the ear (Table 1).

Table 1: Slips of the ear.

Intended	**Perceived**
ENT	*E an**d** T*
Finn	*frien**d****
Creek Inn was	*creek en**d** was**
ham bone	*han**d** bone*

Raising of [ɛ] to [ɪ] before nasals is common in Southern Ohio, where much of this research was done. Therefore, *Finn*, *friend*, *inn*, and *end* all have the vowel [ɪ] for most speakers from this area.

2.2. Other Examples from Slips of the Ear

These examples show that templates developed from hearing equivalent phonological forms in casual speech are used to guide perception and access to meaning.

1. Casual speech phonology allows lexical word-final /t/ to be absent phonetically, as in ‘right person’ pronounced [ˈɹaɪpɚsən]. Here the hearer replaces ‘t’ where it has not disappeared, as it was never there in the first place.

Coke and a Danish	*coconut Danish*

   Again, the ‘t’ is introduced in word-final position, suggesting a larger phonological template.
2. A similar process allows the mapping of certain structures without a phonetic [d] into ones with an underlying /d/, as in ‘hard ball’ pronounced [ˈhɑɹbɔl].
   The perceivers erroneously assume that a final /d/ should be there, so reconstruct it in their percept:

Dierker	*Diergoo**d***
news	*snooze**d***
myofunctional	*mil**d** functional*

   Note that in the last example, the word boundary is mis-assigned, providing evidence that the listener ‘knows’ that apparent deletion of [d] is most common word-finally, again suggesting a misused phonological template.
3. Similarly, ‘st’ clusters can be pronounced as ‘s’, as in ‘first place’ pronounced [ˈfɚsplɛɪs].
   If mis-mapped:

honors political science honest ..	
Goes, like	*ghostlike (here the percept is encouraged by final devoicing)*

4. A sonorant segment takes on the syllabicity of schwa, which does not then appear as a separate vowel, e.g. pronouncing ‘police’ as [pl̩is]
   In these cases, the hearer interprets an ordinary consonant as syllabic:

the urn is finished	*the urine is ...*
Dec writer	*decorator*
fiscal	*physical*
horse story	*horror story*

   In the ‘Dec writer’ example, the listener is oblivious to the word boundary and simply reports a phonological sequence.
5. Another commonly found feature of casual English is weakening of closure, for example, ‘leaking’ is pronounced [lixɪŋ]), with a fricative instead of a stop. Here are some cases of misperceptions where the hearer assumes the consonant has been weakend and erroneously strengthens it:

savor	*sabre*
diverse	*divert*
noon	*nude*
felicity conditions	*ballistic ...*
floor of the house	*Florida house*

6. Final voiced obstruents are completely or partially devoiced, as in pronouncing ‘buzz’ as [bʌs] but (probably) with the appropriate vowel and consonant length for voicing. Here again, we see inappropriate restoration of the voicing. The mis-application requires knowledge that the consonant is final, implying a larger template:

Maple Leaf Weiner	*make-believe*
worse than that	*where’s Annette*

science	*signs*
house plants	*house plans*
parachute	*pair of shoes*
tent pole	*adpole*
slant board	*sled board*

There are, of course, many other kinds of slips of the ear [2] which cannot be related to the phonology of casual speech reductions, though examining them does ultimately contribute to our understanding of speech and language perception. These slips involve all aspects of linguistic structure, including phonetics, phonology, morphology and syntax. Describing these errors is beyond the scope of our presentation, but supports our claim that templates of all sizes are used in perception.

3. Gating

Gating gives even more striking evidence of a holistic perceptual soundscape. It is an experimental paradigm by which a small amount of the acoustic information at the beginning of an utterance is revealed, then another small amount, until the entire utterance is presented. Listeners are asked to make judgments about what they hear as each 'gate' is opened.

In an early experiment [3], we recorded and gated a sentence containing two notable divergences from careful pronunciation. The sentence was 'The screen play didn't resemble the book at all'.

The 'n' at the end of 'screen' was pronounced 'm' (so the word was, phonetically, 'scream') and the word 'didn't' was pronounced [d͂ɪdn], where the second 'd' was a passing, short closure before a nasal release and the final 't' did not appear at all. The gates began in the middle of the word 'screen' and were of approximately 50 msec. At first, all subjects heard 'screen' as 'scream' which is altogether unsurprising, as that is what was said ... there was genuine ambiguity. As soon as the conditioning factor for the n → m assimilation appeared, however, some subjects immediately shifted from 'scream' to 'screen' without taking into account the identity of the following word. This seems good evidence of an active process which resolves the ambiguity, in this case in the correct direction.

Other subjects waited until the end of the word 'play' to institute the reversal of 'm' to 'n' but most had achieved the reversal by the beginning of 'didn't'. Subjects who wait longer and gather more corroborating evidence from lexical identity and/or syntactic structures clearly applied larger templates. With the word 'didn't' the results reflect such a holistic judgement: the word is much more highly-reduced than 'screen' and the time span over which it was recognized was much greater. Three subjects did not identify the word correctly until after the word 'book', and only one subject recognized the word within its own time span. The subjects who did not arrive at a correct interpretation of the entire sentence were those who did not apply the global technique: they arrived at an incorrect interpretation early on and did not update their guess based on subsequent information. Results of this experiment thus suggested that there is a class of very simple phonological processes which can be template-matched quickly, but that processes which seriously alter the structure of a word need to be resolved using a larger pattern.

This may be part of the reason that when asked to identify normal conversational speech in gated form, hearers are often unable to report anything sensible for quite a long time, as can be seen from the following example, which is included to demonstrate what happens perceptually when several casual speech sortcuts are included in the same utterance.

In a similar experiment we gated a highly reduced sentence which was taken from a recorded monologue by a woman in the 30–40 age range about her brother's wedding, "So it was quite good fun, actually, on the wedding, though."

This sentence was chosen for three main reasons: (1) it was one of the few from the recordings of connected speech we had collected which seemed clearly understandable out of context, (2) it contained familiar casual speech reductions (outlined below), presumably having as a basis:

[səʊ ɪt wəz ˈkwaɪt ɡʊd ˈfʌn ˌæktʃʊəlɪ ɒn ðə ˈwɛdɪŋ ðəʊ]

and (3) it had a slightly unusual construction with the major information coming quite late in the sentence. This meant that the well-known phenomenon of words being more predictable as the sentence unfolds was minimized.

We used 30 msec. gates from very near the beginning, and presented the result, interspersed with suitable pauses, to a group of users of Southern British. Scores on perception of the sentence were not perfect: mistakes took place at the very-much-reduced beginning of the sentence, as seen from the following examples of answer sequences:

Subject A
1 i
2 pee
3 pquo
4 pisquoi
5 pisquoi
6 pisquoit
7 ?
8 pisquoifana
9 pisquoifanat
10 pisquoifanactually
11 etc. ... along the same lines)
20 He's quite good fun, actually, on the wedding day.

Subject B
1 tu
2 tut
3 uka
4 uzka
5 she's quite

6 she's quite a
7 she's quite a fun
8 she's quite a fun ac
9 she's quite good fun, ac
10 so it was quite good fun, actually . . .

Notice that the first word in the sentence ('so') was not heard at all by Subject A and quite late by Subject B.

We are far from the first to notice that words are sometimes recognised considerably after they are spoken [4] in casual speech, but most have not focused on the relationship of this fact to phonology. In our experiment, individual sounds may be perceived, but the structure appears only after enough speech to get a correct perspective is achieved, and this can comprise whole phrases. An interpretation arrives very quickly after a satisfactory perspective is achieved, and is rather hard to dislodge, even if it is not correct. Clearly this is a process of assessing the soundscapte before resolving phonetic/phonological ambiguities.

While face recognition is so basic to human perception that some consider it innate, recognition of reduced forms is learned, and learned perfectly by native speakers of a language. Non-native speakers of English generally show very poor performance in decoding gated conversational English, probably because they are not briefed to expect shortcuts. But perhaps not surprisingly, not all non-native speakers perform in the same way.

To test this, we ran another experiment in which the same gated sentence described above was used as stimulus, All of the suprasegmental cues for position in utterance were present, including the intonation pattern.

Phonetically, the sentence was:

[səʷɪʷsʷˈkwaɪʔgʊˌfʌnætʃʊiɒn̪ːəˈwɛdɨŋ..d̪ɵʉ]
/səʊ ɪt wɒz.......ækʃʊəli......................ðəʊ/

There was no 't' in 'it', the [w] in 'was' was represented by rounding in the first syllable, the 't' in 'quite' was a glottal stop, there was no 'd' in 'good', 'actually' was quite reduced, there was no separate dental fricative in 'the', and the fricative at the beginning of 'though' was pronounced as a dental stop.

The utterance was presented in a gated fashion (20 gates of 30 msec, with ten seconds between each stimulus), and subjects were asked to write what they heard in normal spelling after each stimulus. Two test stimuli were presented before the writing began, to accustom subjects to the input.

The recording was presented to four groups of subjects, five native speakers of English, eight Greeks between the ages of 20 and 30 who had studied in the UK for an average of 3 years, and ten Polish students with the same profile, and 13 Polish students of English who lived in Poland.

4. Results

Results for a random selection of native English speakers are shown in Figure 3. The grey bars show where in the utterance the word at the left was recognised. The number inside the grey bar shows how many subjects made the identification at that time. The column 'Total' on the right gives the total number of correct answers for each word. The pattern of results that you see here is much like other results we have collected: perception follows the time course of the utterance, but not in a strictly sequential manner. Often two or more words are identified at the same time.

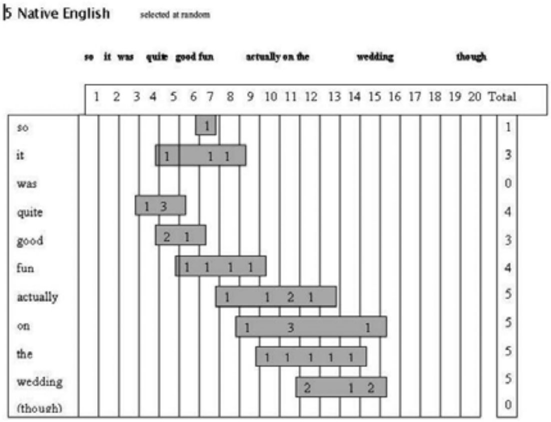

Figure 3: Native English speakers, selected at random.

In this case, only one native speaker reported 'so, and that was after or at the same time as 'it'. No one reported 'was', though a few people wrote 'is'. The following seven words were all relatively well identified, with some sequentiality, but there was a strong tendency to hear more than one word at a time, as represented by the 'stacks' of grey boxes. No one heard 'though'; rather, everyone reported 'day': it is clear from the phonetic transcription that the vowels were quite fronted, and this together with the expectation of the familiar phrase 'wedding day' led to a wrong conclusion. Total correct score was only 64% overall, though the score reached 85% on the next seven words.

In Figure 4 we see results when the same gated utterance is presented to Greek graduate students at the University of Reading.

The Greek subjects predictably showed less successful perception than the native speakers, averaging 25% correct overall. Like the native speakers, the poorest results were for the first three words and the best results for the word 'quite'. On the subsequent seven words, their score was 34%. One speaker did hear 'though', in this case excelling the native speakers.

Results are suprisingly different for a group of Poles with a similar educational profile.

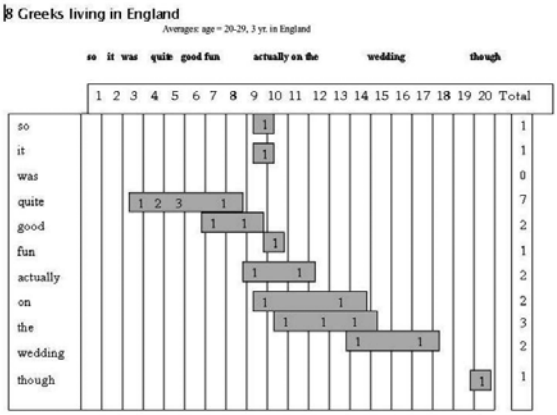

Figure 4: Greek young adults studying in England.

4.1. Polish young adults studying in England

Some of the profile, shown in Figure 5, is similar to that of the Greeks: the first three words were poorly identified and the following ‘quite’ universally recognised. But the overall score was 51%, with 71% on the next seven words. Four Poles heard ‘though’, excelling the native speakers.

It is interesting that though there is some evidence of simultaneous recognition of words, in general word recognition follows the time course of the input, a feature not so obvious for native speakers. This suggests that while the Poles are doing some holistic processing, they are in general finding smaller patterns than native speakers of English.

It is tempting to conclude that Poles may have an advantage when learning English, and this is partially borne out by results from 11 Polish students of English living in Poland.

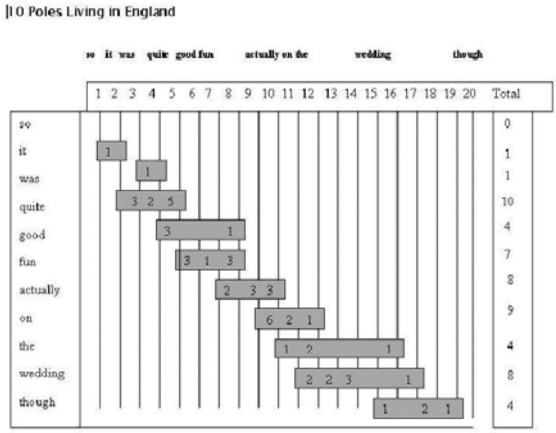

Figure 5: Poles living in England.

4.2. Polish students of English living in Poland

Figure 6 shows the results for Polish students of English living in Poland. Here the 'totals' column reflects students at two different levels of study. The first number is the total for all subjects, the first number in parentheses is results from 5th-year students, and the second is results from 7th-year students[1].

In this case, no one identified the first three (very highly reduced) words; 'quite' and 'wedding' (the two most stressed words) got high scores, but 'good' and 'fun' did not fare well. Three subjects heard 'though', much the same as the Poles living in England. Here, again, we see 'stacking' of results, showing that several words were identified at nearly the same time.

The overall accuracy rate was 36% and the score on the most-recognised seven words, 48%.

While the last two sets of results clearly indicate that living in the English language environment is very beneficial for the understanding of casually-spoken English, it is still striking that the Poles living in Poland scored higher than the Greeks who had been living in England for three years.

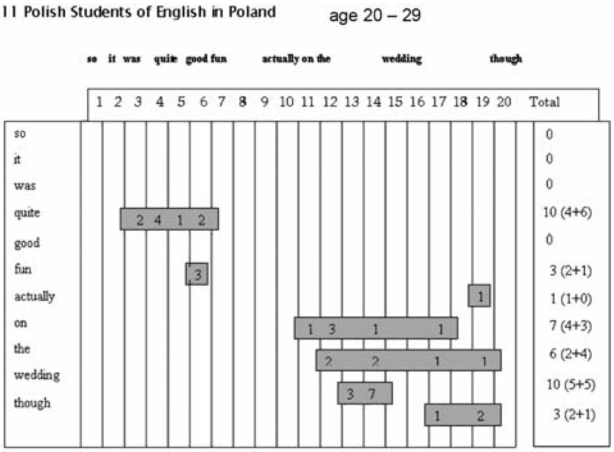

Figure 6: Polish students of English living in Poland.

We are left asking why Polish students appear to have an advantage in identifying English casual speech intentions, and a number of guesses can be offered, including excellent teaching and high motivation. A possible reason is that Poles have similar syllabic patterns in their own language, while Greek syllable structure is entirely different and, on the whole, simpler.

Polish and English are unusual in that they allow several consonants at the beginning and at the end of a syllable. It is well known that the most usual syllable structure in the world's languages allows at most one consonant at the beginning of a syllable and none at the end (the CV syllable), and while Greek has interesting initial clusters, syllable-final consonants are highly constrained. Clusters in English and Polish are different

[1] With many thanks to Professor Wiktor Gonet, who collected the data for us.

phonetically, but they share complexity. Poles appear to be more able than Greeks to predict what kinds of shortcuts will be taken. Whether they have similar reduction patterns in their own language calls for investigation.

5. Conclusion

Following Natural Phonology (Stampe [5–7]), we assume that phonological simplifications are governed by the same principles which shape phonological sequences in all languages.

The study of casual speech phonology across languages is not well advanced, and the potential for generalisation is small at this stage. In future work, we hope to investigate the relationship between articulatory complexity and casual speech shortcuts in other languages and to relate this to perceptual strategies. We hope that a comprehensive theory can eventually provide implicational algorithms such as ‘if a language has syllables of shape X, it will show simplifications of type Y’, and that, as above, these will be shown to provide perceptual templates for speakers of that language.

References

[1] Bard, E. G., R. C. Shillcock and G. T. M. Altmann. 1988. The recognition of words after their acoustic offsets in spontaneous speech: effects of subsequent context. *Perception and Psychophysics* 44, 395–408.

[2] Bond, Z. 1999. *Slips of the Ear: Errors in the Perception of Casual Conversation*. Academic Press.

[3] Shockey, L. 2003. *Sound Patterns of Spoken English*. Blackwell.

[4] Shockey, L. and Watkins, A. 1995. Reconstruction of base forms in perception of casual speech. In K. Elenius and P. Branderud, eds., *Proceedings of the 13th International Congress of Phonetic Sciences*. Stockholm. 588–91.

[5] Stampe, D. (1973) On chapter nine. In M. Kenstowicz and C. W. Kisseberth, eds, *Issues in Phonological Theory*. The Hague: Mouton. 44–52.

[6] Stampe, D. (1979) *A Dissertation on Natural Phonology*. Chicago University Press.

[7] Stampe, D. (1987) On phonological representations. In W. U. Dressler, H. C. Luschusky, O. E. Pfeiffer and J. R. Rennison. eds., *Phonologica 1984*. Cambridge University Press, 287–99.