INVARIANT AUDITORY WORD PATTERNS IN SPEECH PROCESSING

NATALIE WATERSON

Department of Phonetics and Linguistics
School of Oriental and African Studies
University of London, London, England

ABSTRACT

It is proposed that one of the units of speech perception is an invariant auditory word pattern. This consists not of the whole spectrum but of a limited number of acoustic cues that are auditorily salient, together with those that are less salient but carry contrastive function in the language. Speech processing takes place by pattern recognition and pattern matching. For this two levels of representation are postulated, a phonetic level, LR1, and a lexical-phonological level, LR2. Cues are abstracted from the acoustic signal and are synthesized into patterns: these are checked against patterns at LR1. If they match, they are then matched with patterns at LR2 and indentification of the word is achieved. The organization of patterns in a network is shown for a sample of a child's phonological system, and how recognition of some words takes place is illustrated. An example of a misperception is also given to show how confusions occur between words of same patterns.

It is known that there is much redundancy in speech and that speech processing is very rapid. Context, knowledge of the language, knowledge of the topic, shared knowledge, etc., are acknowledged to play a major role in the interpretation of speech. Because speech processing is so rapid, it is clear that interpretation of the acoustic signal segment by segment is not possible. Furthermore, no one-to-one acoustic correlates have been found for phoneme segments. Word by word processing is still too slow to explain the speed with which speech is processed. It is possible that the auditory processing of speech is similar in nature to visual processing in the interpretation of written texts by reading. It is recognized that in reading attention is not given to each letter of each word, nor even to each word, and that scanning takes place—not in a serial progression but with the eyes moving back and forth as they abstract the most essential information from the text to make sense of the message. Words have visual shapes which aid recognition. It seems that one may similarly look for auditory shapes of words that can be recognized in auditory scanning of the acoustic signal. There may similarly be auditory shapes of sentences where attention is not given to the whole of the acoustic signal but abstractions are made at points (stressed high information words) which will give maximum information for the minimum expenditure of time and energy. It is proposed that it is auditory word patterns that are abstracted at such points in the acoustic signal and that such auditory patterns are invariant.

Auditory patterns are essentially acoustic skeletons composed of auditorily salient cues and such less salient cues as have contrastive function in the language. Thus the pattern will consist of part of the spectrum of the word, not the whole spectrum. The cues will involve mainly features such as intensity, e.g. peaks will indicate the number of syllables, greater intensity of some peaks and lesser of others will mark strong and weak stress; duration will mark some syllables longer than others; fundamental frequency will indicate the pitch pattern; first formant will indicate the various degrees of openness of vowel, i.e. whether more close, more open, or the same as in adjacent syllables; cues for different classes of consonants will differentiate them from each other, e.g. a fricative from a nasal, a nasal from a plosive, etc., etc. Such cues in different sequences comprise the different word patterns. The cues are relative, i.e. it is not the actual intensities, actual frequencies, actual durations, etc., that are relevant, but the relationships between the cues, which are fixed. The patterns are therefore invariant and remain the same regardless of variables such as if speech is compressed, as in fast tempo, or whether spoken at a very slow tempo; whether spoken on a man's low pitch or a woman's high pitch, and whether pronounced in the standard dialect or some provincial dialect. Words in the phonological system of a language may be described in terms of invariant auditory patterns. There may be several words belonging to a single pattern or only one or two. For instance the words: train, plane, prim, cream, tram, clan, may be classed as belonging to the same pattern, plosive with fricative release + vowel + nasal, PFVN. The fricative release may be lateral or central, and the vowel may be long or short, open, close, or mid with glide to close. Words like pot, kick, deep, bat, boot, belong to another pattern, plosive + vowel + plosive, PVP, and so forth.

Patterns will be organized for rapid and easy retrieval in a network which has two levels of representation (leaving aside at present the semantic and syntactic levels) the first being phonetic for receiving the acoustic signal and for synthesizing patterns and for storing patterns; the second level is for storing the phonological patterns for matching with the phonetic patterns, to arrive at the identification of words.

The adult processing system, with a vocabulary of many thousands of words is extremely complex, so for illustration, a sample of a child's very early, very simple phonological system will be used to demonstrate the proposed network and how recognition takes place. The child, aged between 1;5 and 1;6, had monosyllables and disyllables involving mostly nasals and plosives. Fig. 1 shows the network of LR1 and is mainly phonetic; Fig. 2 shows LR2 which is lexical-phonological. How such levels of representation are constructed by the child and how processing takes place in relation to these levels of representation is described in detail in [1]. In the construction of LR1, patterns are synthesized on the basis of auditorily salient features of the acoustic level. A child's limited abilities, especially in perceptual discrimination, oblige him to pay selective attention to what is most acoustically and auditorily salient at first. The patterns are stored at LR1 for future matching when other words of such patterns are recognized. Patterns of LR1 are more fully specified at LR2, and meaning is included. LR1 patterns are matched with patterns at LR2 in the process of recognition. If there is no match for the synthesized pattern, a new pattern is constructed and stored.

Fig. 1 shows the monosyllablic and disyllabic patterns of the child's LR1 and Fig. 2 shows the organization of one monosyllabic pattern, the PV pattern, in LR2. The way recognition takes place is by following pathways along which choices are made which constrain the possibilities and lead to the identification of the particular word. The pathways for the words are shown by different markings (see key on figures). It will be seen that the PV pattern words follow the same path up to the point where they divide according to the degree of vowel openness, marked by ɑ for low vowel, ɛ for mid and ɪ for high vowel. The next choice is at the three-way contrast carried by place of articulation, viz. labial p, apical t, and dorsal k. The last choice is of contrasts carried by frontness y, backness w, and neutrality as to frontness and backness, ə, and the pattern is then identified as the particular word.

In the case of a child, the early forms are based mainly on the auditorily salient features of words which are fleshed out within his current capabilities and in a way that fits his current network. As he becomes able to give more attention to less salient features, his forms of words and patterns change and his network is therefore constantly being re-structured. For instance when [bəu] 'boat' acquires a final plosive, it moves from the PV pattern to the PVP pattern, and when [gu:] 'goose' acquires a final fricative, [gu:θ] and [gu:ɸ], a new pattern has to be created and incorporated into the network, viz. PVF (plosive + vowel + fricative) to which will belong newly acquired words like [bɪf] 'beef' and [gauf] 'cow' and 'calf'. Eventually the child acquires the complex network of the adult. This concept of invariant auditory pattern can thus offer an explanation of how the acquisition of phonology takes place.

Further evidence in support of the invariant auditory word pattern can be found in studies of misperceptions (see [1]; also for references for support from other disciplines). Examples show how much the listener contributes to the interpretation from what he thinks the intended message is, making use of the minimum of acoustic cues of the pattern and the maximum use of any other available information. A brief illustration is given of the way the interpretation of an utterance is made in terms of pattern recognition, together with use of context, shared knowledge, and other factors. It will be shown how non-linguistic information influences the interpretation of a pattern which results in the identification of the wrong word.

Context: Saturday morning. A and B are in the bathroom and the bath is being filled with water, so there is a loud noise of rushing water which has a masking effect. A and B had just been talking about changing the positions of their parked cars to enable A to take C to the station to get the 8.48 train. The following conversation then takes place.

B: We must get the E.45 cream today.
A: Today? Why today?
B: Why not?
A: Why on a Saturday?
B (Realizing that A has got the message wrong): I said 'We must get the E.45 cream today.'
A: Oh, I thought you said 'We must get the 3.45 train today.'

A was still geared to the semantic field of trains and train times and did not realize the change of topic, and as B and her husband often came for weekends, arriving on Friday night and usually leaving on Sunday, A misinterpreted the pattern common to 'cream' and 'train', viz. plosive with fricative release + vowel + nasal, PFVN, as 'train'. She also interpreted 'E' [i:] as 'three' [θri:]. Voiceless non-salient [θr] would easily be masked by noise and a listener would therefore be ready to 'restore' it where needed, as here, where '[i:] forty-five' could only mean '3.45' in terms of train times. A having recognized the pattern PFVN, it is possible that detailed pattern matching would be skipped as the context so clearly predicted 'train'. In fact, B was referring back to the previous day's conversation (shared knowledge) about a cream called E.45 which she had recommended to A.

This example shows how the processing of the invariant auditory word pattern in combination with the use of non-acoustic information can speed up the rate of speech processing. Because of adults' huge vocabularies and complex phonetic, phonological, semantic and syntactic systems, and their fast rate of speaking, adults need to use the maximum possible short cuts in processing. The concept of invariant auditory word pattern makes it possible to explain how short cuts in processing can be made and why speech processing can be as rapid as it is.

Intonation patterns have long been described as a limited number of invariant tunes and the problem of normalization across variables as a man's use and a child's use of the tunes does not arise. Similarly, the proposed auditory word pattern is also invariant and the problem of normalization across variables such as age, sex, speech rate, and dialect need no longer arise.

[1] N. Waterson, «Prosodic Phonology: The Theory and Its Application to Language Acquisition and Speech Processing», Grevatt & Grevatt, 1987.

Figure 1

Key: C = consonant system
V = vowel system
P = plosive system
N = nasal system
S = sibilant system
* indicates continuity of pathways between figures 1 and 2

broken lines, dotted lines, etc.: see **Key** in Fig. 2
(see Fig. 2 for continuation of pathways at PV)

Fig. 1: **Organization of patterns at LR1**

Figure 2

9 [bæ] [bɑ:] [ba:] [dau] [gɑ:] [gau] [bə:] [bəu] [bɔ:] [bʌ] [dɔ:] [dɛə] [dɔ] [bu] [du:] [di:] [gu:]
bag boy Bob down car cow bird boat ball letter 'b' door chair dog bull two cheese goose

Key: * see **Key** in Fig. 1
α = low vowel system
ɛ = mid vowel system
ι = high vowel system
— over vowel = long length
˘ over vowel = short length
p = labial term of C system
t = apical term of C system
k = dorsal term of C system
y = prosodic frontness
w = prosodic backness
ə = neutral as to y and w

pathways for recognition of: car....boat----cheese→→
two._._._Bob—·—·
(see Fig. 1 for pathways before PV)

Fig. 2: **Organization of PV pattern words at LR2**