ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

PHONOLOGICAL PROCESSING IN THE AUDITORY SYSTEM: A NEW CLASS OF STIMULI AND ADVANCES IN FMRI TECHNIQUES

Roy D. Patterson, Stefan Uppenkamp*, Dennis Norris**, William Marslen-Wilson**, Ingrid Johnsrude**, and Emma Williams****

*Centre for the Neural Basis of Hearing, Department of Physiology, University of Cambridge, Downing Street, Cambridge, CB2 3EG, U.K.
**MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 2EF, U.K.
***Wolfson Brain Imaging Centre, University of Cambridge, Box 65 Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, U.K.

ABSTRACT

It is commonly assumed that, in the cochlea and the brainstem, the auditory system processes speech sounds without differentiating them from any other sounds. At some stage, however, it must treat speech and non-speech sounds differently. In broad terms, the purpose of this paper is to consider where this speech specific processing begins in the auditory pathway. Specifically, the paper is concerned with extrapolating the concepts of an auditory model to the point where we can define matched sets of speech and non-speech sounds that can be used in a brain-imaging experiment to delimit where phonological processing of vowel sounds begins in the auditory system. Pilot results suggest that phonological processing of vowels may begin just outside auditory cortex in Brodmann area 21.

1. THE TRANSIENTS, TONES AND NOISES OF SPEECH

Sounds in the natural world fall broadly into three categories: **transients, tones and noises**. Hearing research suggests that the auditory system constructs internal ***auditory images*** of sounds [1], and that the images of transients, tones and noises exhibit large, characteristic differences [2]. Humans produce all three categories of sounds when they speak, and the concept of the auditory image can be illustrated using components of the word 'past'. From the auditory perspective, the word consists of a transient puff of air (the plosive consonant, /p/), a tone in the form of a stream of glottal pulses (the vowel, /ae/), a burst of broadband noise (the fricative consonant, /s/), and another pulse of air (the stop consonant, /t/). Auditory images of the vowel, /ae/, and the fricative, /s/, are presented in Figure 1. There are 60 channels in the cochlea simulation that performs the spectral analysis; the image consists of 60 time-interval histograms, one per row of the auditory image. The ordinate of the image is the centre frequency of the auditory filter used to construct the channel, and so the ordinate corresponds to the tonotopic dimension in the cochlea. The auditory image of the noise, /s/, shows activity at all time intervals across the image but there is no regularity or structure in the image. The absence of pattern is the characteristic of noisy sounds, both on the macro and micro scale. The regularity about the 0-ms vertical is an artefact of the time-interval calculation; the intervals are calculated from peaks in the motion of the cochlear partition and the peak at the start of the interval is always mapped to 0-ms in the auditory image. The average level in the image is fixed for a stationary noise, but the level varies from moment to moment and it is this which corresponds to the hiss in the noise.

The auditory image of the tone, /ae/, reveals the cochlea's response to glottal pulses; since they repeat regularly, there is activity at time intervals corresponding to multiples of the glottal period (~8 ms). The presence of a repeating structure across the image is the characteristic of tones in the natural world. The *horizontal spacing* of the auditory figures reveals the *pitch* of the sound; the spacing between auditory figures decreases as pitch increases. The lower limit of pitch is about 30 Hz, corresponding to a period of 33 ms [3], and this is the maximum width of the auditory image.

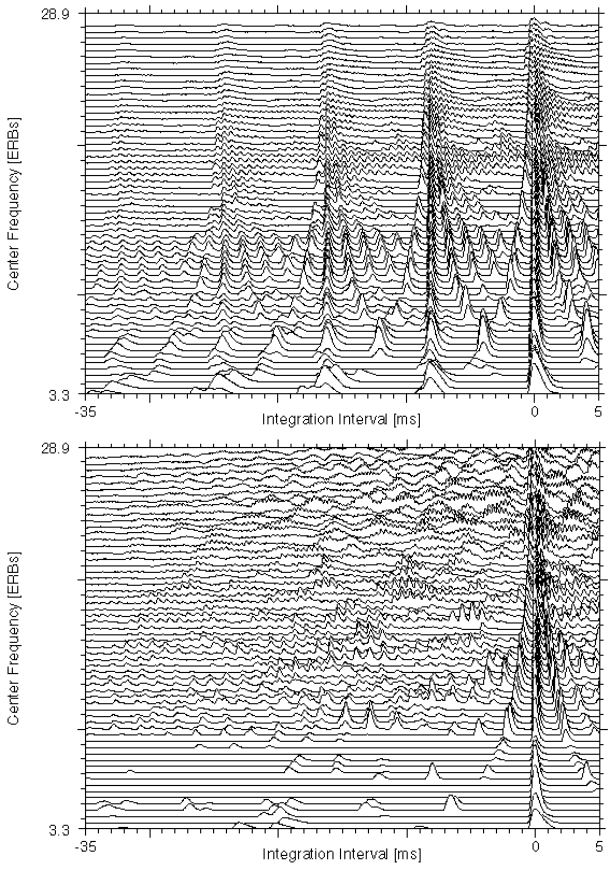

Figure 1: Auditory images of a tone, the /ae/ in 'past' (top) and a noise, the /s/ in 'past'(bottom).

The formants of the vowel appear as concentrations of activity at points on the vertical structure. The structure itself is referred to as an ***auditory figure*** and so vowel quality corresponds to

the shape of the auditory figure. The auditory figures produced by a tonal sound do not move within the image so long as the pitch is stationary. When the rate of pulses is high relative to the decay rate of the image, as it is for vowels, the image is stable in level for the duration of the stationary part of the sound. Moreover, for vowels the rate of change of pitch is slow relative to the rate of glottal pulses and so the image expands or contracts smoothly as the pitch falls and rises, and the formants move smoothly up and down the auditory figure as vowel quality changes. So, pattern, temporal regularity, temporal stability and smooth motion are the characteristics of tonal sounds in the auditory image.

The auditory image of a transient, like the /p/ or /t/ of 'past' consists of one auditory figure centred over the 0-ms point in the image, with essentially no activity in the remainder of the image. It is essentially the neural version of the multi-channel impulse-response produce by a click in the cochlea. The half life of the auditory image is just 30 ms and so transients only appear for a moment in the image.

2. MATCHED SPEECH AND NON-SPEECH SOUNDS

Brain imaging studies in speech research are typically more concerned with locating the centres involved in lexical or semantic processing rather than phonological processing; that is, more concerned with the later, rather than the earlier, stages of speech processing. Accordingly, they use continuous speech and contrast it with, for example, rotated or reversed speech [4] to preserve and balance the spectral and temporal complexity of the sounds in their imaging contrasts. Transients, tones and noises are not transmuted from one category to another by reversal or rotation; they remain transients, tones or noises, albeit slightly different instantiations of these sound types. So any centre involved in processing transients, tones and noises as such will be just as active for reversed and rotated speech as for normal speech, and not surprisingly, all of these speech and non-speech stimuli activate the larger part of the temporal lobe when contrasted with silence. To locate where phonological processing begins, we need a more specific contrast, and preferably one that does not involve lexical, syntactic or semantic processing. Accordingly, we have developed a new class of synthetic vowel sounds to try and delimit the point in the auditory system where phonological processing begins.

The basic building block of the synthetic vowels is a 'damped' sinusoid [5] constructed by applying an exponentially decaying envelope with 4 ms half-life to a short segment of a sinusoid (Figure 2). A single damped sinusoid is like one cycle of a formant in a vowel. The upper left panel shows four damped sinusoids with the same 16-ms envelopes. The carrier frequencies are fixed at the formant frequencies of /a/, and the sound produced by the sum of the damped sinusoids (bottom row) automatically activates the phonological system and produces a speech perception provided the sound is syllable length (about 300-400 ms). The remaining three panels show how we can produce control sounds with very similar distributions of energy over frequency and time, some of which activate the phonological system and some of which do not.

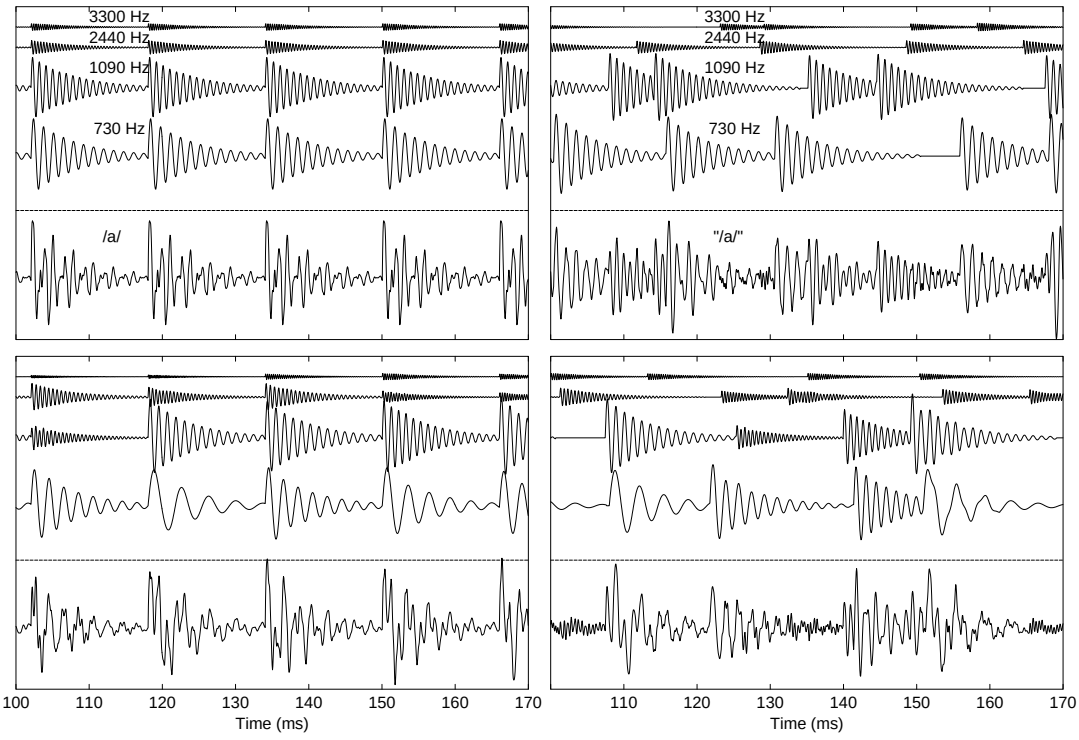

Figure 2: Four classes of stimuli constructed from sets of isolated formants (damped sinusoids). The resulting stimuli all have similar long-term distributions of energy over frequency and time. In the upper row, the carrier frequencies are fixed and the stimuli are heard as vowels. In the lower row, the carrier frequencies vary and the stimuli are heard as 'musical rain'.

In the upper right panel we have randomised the start points of the damped sinusoids within each 16-ms 'cycle' while keeping the carrier frequencies fixed at formant frequencies. The resulting sound is still recognisable as an /a/ but from a vocal tract with a pathalogical degree of jitter. In the lower left panel, we have randomised the carrier frequencies within their formant bands keeping the start points fixed, and in the lower right panel both the start points and carrier frequencies have been randomised. These sounds do not activate the phonological system at all; indeed, they sound like two strange forms of 'musical rain', one with a continuous low pitch due to the synchronous start points. They do, however, have the same average statistics as those in the upper panels, and so will serve as the appropriate controls in the search for the location of phonology and pitch centres in the auditory system.

3. PERCEPTUAL QUALITY OF SYNTHESISED SOUNDS

A paired comparison experiment was performed to quantify the perceptual quality of the synthetic vowels relative to the non-vowel sounds. Eighteen sound conditions were presented in this experiment, including the four sounds shown in Figure 2. Table I summarises the sound conditions; the sounds from Figure 2 are marked "*". The sounds differed in the amount of randomisation of the position of the damped sinusoids both in the time and frequency domains. The procedure was a two-interval, two-alternative forced choice task. Each stimulus interval contained a sequence of three randomly chosen stimuli (vowels or non-vowels) from one particular sound condition. The stimuli had a duration of 400 ms and they were separated by 200 ms of silence. The two stimulus intervals within one trial were separated by 500 ms and their onsets were marked by lights. The listeners were asked to choose the interval that sounded most vowel-like. In the case of two completely non-vowel-like sounds, they were asked to choose the one that was more like speech. They could repeat a trial once before they were forced to give a response. All stimuli were scaled to have the same RMS level. They were played by a TDT system II at a sampling frequency of 20 kHz into a lowpass filter with cutoff at 8 kHz, and presented diotically via headphones (AKG 240D) at 50 dB HL.

Nine normal hearing subjects participated in the experiment. Before the main experiment with all sound conditions, several examples were played to illustrate the full range of the stimuli. During the experiment, no stimulus was compared to itself, and each comparison was carried out twice, once with the order A-B, and once with the order B-A, and so there were $18 \cdot 17 = 306$ trials. The order of the trials was randomised, and the experiment was divided into 17 runs with 18 trials. The subjects were asked to take a short break about every 4 runs. A complete session for one subject lasted for approximately one hour.

A relative psychophysical scale of preference, reflecting the speech-like quality of the sounds, was constructed from the judgements using the Bradley-Terry-Luce method [6]. This method is based on a linear model which assumes that on the dimension of interest, the stimuli can be ordered according to a linear scale. Pooling the judgements from all nine listeners gives a total of 2754 trials, or 18 observations for each pair of stimuli (9 A-B, 9 B-A) and 153 observations for each single stimulus. Figure 3 shows the resulting hierarchy of preference. Arrows indicate the positions of the sounds from Figure 2. The whole scale covers a range of approximately 5 points (-2.5 to 2.5). This scale should be read as a relative scale; that is, only differences have meaning. The zero-line is intrinsic to the analysis and the value has no particular meaning with regard to perceptual quality.

With regard to the two dimensions of randomisation, carrier frequency and onset time, it is obvious that formant frequency is essential for producing a vowel perception. The five sounds at the top of the scale have fixed, proper, formant frequencies, while four of the five sounds at the bottom end of the scale have randomised carrier frequencies. Speech pitch, produced by regularising the onsets, is the next most important property for producing a vowel percept. Two-formant, regular, damped vowels are preferred over "pathological" four-formant vowels (no pitch), and they are rated as much more speech-like than four-formant sinusoidal vowels without damped envelopes (sin_vow). Adding the damped envelope, either with periodic or with random timing, results in an increase of nearly 1.5 points on the preference scale. Even regular, single-formant, damped sounds are rated as more speech-like than any sound made out of flat-envelope sinusoids.

dmp_vow	* damped vowels, four tracks of damped sinusoids at formant frequencies
dmp_two	as dmp_vow, but only first and second formant
dmp_fst	as dmp_vow, but first formant only
dmp_snd	as dmp_vow, but second formant only
flt_vow	as dmp_vow, but no lowpass slope in spectrum
jit_vow	as dmp_vow, 10% jitter in envelope timing
pth_vow	as dmp_vow, irregular envelopes (100% jitter in timing), i.e. no pitch
noi_vow	* as dmp_vow, but narrow bands of noise as carriers of triangles
sin_vow	four sinusoids at formant frequencies, no damped envelope
sin_two	as sin_vow, but only first and second formant
sin_fst	as sin_vow, but first formant only
sin_snd	as sin_vow, but second formant only
noi_pit	four tracks of damped noise bands, one octave wide
noi_ran	as above, but irregular envelope (no pitch)
fxr_pit	four tracks of damped sinusoids, random change of carrier frequencies within limited bandwidth, regular timing
fxr_ran	as above, but random timing (no pitch)
mus_pit	* complete randomisation of carrier frequencies for each track, regular timing
mus_ran	* randomisation of carrier frequencies and timing

Table I: Sound conditions used in the paired comparisons experiment.

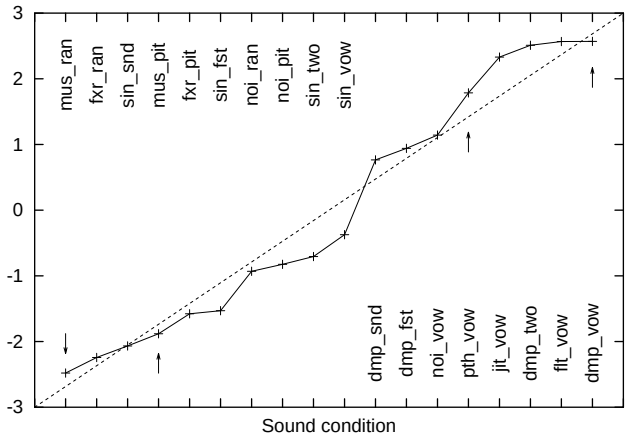

Figure 3: Relative scale of preference for synthetic vowels and non-vowel sounds from a paired-comparison experiment.

On the other end of the scale, randomisation of both the carrier frequency and the onset time within each track produces sounds which sound very different from human vowels; they are more like some kind of "musical rain". Synthesising the vowels as sets of single damped sinusoids centred at formant frequencies, and with a realistic repetition rate, enables the creation of sounds covering a wide range of perceptual quality from speech-like to completely non-speech like by means of simple manipulations in the time and frequency domains. Damped vowels and "musical rain" have similar long-term distributions of energy in time and frequency, so they should activate primary auditory areas in a very similar way. Any area in the brain that extracts phonological information from incoming sound should show a strong contrast between these two conditions.

4. ACTIVATION OF A CANDIDATE REGION FOR PHONOLOGICAL PROCESSING

Until recently, the use of fMRI in hearing research was severely limited by i) scanner noise that interferes with the stimuli and reduces sensitivity to the response, and ii) the need to present sounds down long plastic tubes to avoid having metallic headphones in the scanner. Hall et al. [7] introduced 'sparse imaging' to solve the scanner noise problem: Within each 12-s interval, the scanner is silent for 9 s while the experimental stimuli are presented and then a full volume of slices is gathered in one 3-s burst. They showed that sensitivity is increased because the response to the scanner noise itself is separated from the response to the experimental sounds. Palmer et al. [8] and others developed a magnet friendly headsets to solve the fidelity problem; at the same time, the headsets were mounted in muffs designed to minimises scanner noise reaching the cochlea via the ear canal.

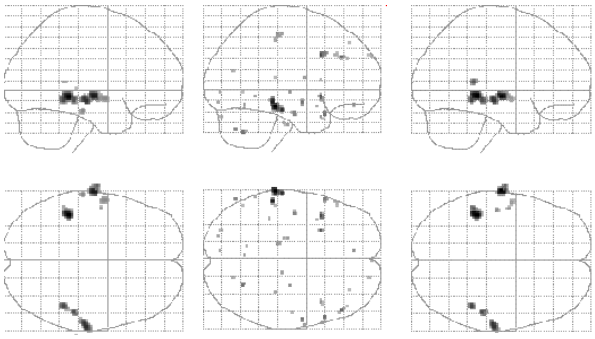

Figure 4: Pairs of sagital (top) and axial glass brains for the contrasts damped vowel vs silence (left), damped vowel vs musical rain (middle), and musical rain vs silence (right).

A pilot experiment was performed with a 3 Tesla scanner and one listener to determine the number of volumes required to produce strong activation with the damped-vowel and musical-rain stimuli, and provide sufficient sensitivity to examine the contrast. It was hypothesised that regions where damped vowels produce more activity than musical rain are candidates for the speech phonology centre. A set of 48 volumes was gathered for both of the experimental stimulus conditions and for a third silence condition as a baseline. The results for the three conditions are presented in Figure 4 as pairs of sagital (top) and axial (bottom) 'glass' brains. When compared with silence, damped vowels (left column) and musical rain (right column) produce remarkably similar patterns of excitation restricted to the surface of the temporal lobe bilaterally. The activation is centred around Heschl's gyrus in the right hemisphere; in the left hemisphere it is more anterior, more lateral and posterior. The similarity of the patterns and the lack of activation outside these regions illustrates the value of the musical-rain control condition. The central panels show that there is one region where the damped vowels produce more activation than the musical rain; it is in the left hemisphere as would be expected, on the lateral surface of the temporal lobe, below and posterior to Heschl's gyrus and the planum temporale.

5. CONCLUSIONS

The auditory images of tones, noises and transients suggest a method for producing matched sets of sounds that either do, or do not, produce vowel perceptions. The perceptual distinction between these damped vowels and musical rain was established experimentally in a large paired comparison experiment. Subsequently, a pilot fMRI study with a high-fidelity, magnet-friendly sound system and a sparse imaging technique revealed a candidate location for phonological processing in the Brodmann area 21, just below and behind auditory cortex.

Research supported by the U.K. Medical Research Council (G9703469).

6. REFERENCES

[1] Patterson, R.D. (1994a). "The sound of a sinusoid: Time-interval models," *J. Acoust. Soc. Am.*, 96, 1419-1428.

[2] Patterson, R.D., Allerhand, M., and Giguere, C., (1995). "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.* 98, 1890-1894.

[3] Krumbholz, K., Patterson, R.D., and Pressnitzer, D. (2000). "The lower limit of pitch as determined by rate discrimination," *J. Acoust. Soc. Am.* (in press, Sept).

[4] Mummery, C., Ashburner, J., Scott, S., and Wise, R. (1999). "Functional neuroimaging of speech perceptionin six normal and two aphasic subjects," *J. Acoust. Soc. Am.*,106, 449-457.

[5] Patterson, R.D. (1994b). "The sound of a sinusoid: Spectral models," *J. Acoust. Soc. Am.*, 96, 1409-1418.

[6] David, H.A. *The method of paired comparisons.* 2nd ed., Oxford University Press, New York, 1988.

[7] Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., Bowtell, R.W. (1999). '" Sparse' temporal sampling in auditory fMRI," *Hum. Brain Mapping*, 7, 213-223.

[8] Palmer, A.R., Bullock, D.C., Chambers, J.D. (1998). "A high-output, high-quality sound system for use in fMRI," *Neuroimage*, 7, S359.