A Model of the Perception of Concurrent Vowels at Short Durations

S.L. McCabe and M.J. Denham
Centre for Neural and Adaptive Systems
School of Computing, University of Plymouth, Plymouth PL4 8AA, UK
suem@soc.plym.ac.uk, mike@soc.plym.ac.uk

Abstract

The perception of double vowel stimuli has previously been modelled on the basis of determining the pitch of the component sounds, using this information to segregate frequency channels which comprise each of the sounds, and then using a 'vowel' template to identify the stimulus. Such models fail to adequately account for the human ability to distinguish vowels with the same fundamental frequency, or the effects of vowel inharmonicity, and have difficulty in explaining the results of experiments using short duration stimuli. The model we describe here is based on the connectivity of the thalamocortical system and shows how a competitive recognition process and top-down suppression of the dominant vowel can lead to the identification of both vowels even in the absence of pitch grouping cues. This model provides a novel explanation for the perception of double vowel stimuli and is consistent with recent results showing the dependence of the phenomenon on stimulus duration.

1. Introduction

Auditory scene analysis addresses the problem of discovering how incoming acoustic signals are decomposed into perceptual representations of sound sources in the environment. Double vowel experiments, in which subjects are presented with stimuli formed by superimposing two vowels, have been used extensively as a means of investigating segregation within the auditory system within a carefully controlled task [1,2,3]. Humans are surprisingly adept at identifying both of a pair of vowels which have common onset, offset, direction, similar loudness and even common fundamental frequency or pitch. Figure 1 illustrates typical human performance in such experiments [2]. As can be seen differences in pitch result in increases in performance of up to 20%, hence most models of this process have been based on the assumption that the auditory systems first determines the pitches of one or both of the vowels and partitions the frequency channels on the basis of common pitch [2,9]. A template matching procedure is then used on each of the channel subsets to determine the constituent vowels. These models achieve reasonable performance in replicating human perception when there are differences in fundamental frequency but fail to adequately account for the perception of vowels with a common fundamental frequency.

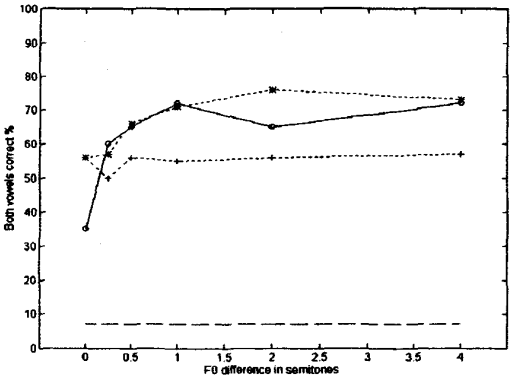

Figure 1. Probabilities of identifying both constituents of a double vowel stimulus correctly. '*' indicates performance for stimulus durations of 200 ms, and '+' for durations of 51.2 ms; '--' indicates chance performance in this task.'o', connected with solid lines, indicates Meddis and Hewitt's model performance. Adapted from [2,9].

The assumption of pitch based segregation, although apparently reasonable, has also been called into question by recent experiments which have shown that listeners can identify double vowel stimuli even in situations when they clearly have no ability to identify

0-7803-4859-1/98 $10.00©1998 IEEE

pitch [8]. The perception of pitch actually seems to require stimuli of fairly long duration [13] and subjects can frequently make vowel judgements even when they cannot determine the pitch of the sound [8,12]. Although the lack of conscious awareness of pitch does not prove that people are not using pitch, it seems likely that this cue may only be useful at longer durations and as shown in figure 1, the improved performance with increasing difference in fundamental frequency is lost with short duration stimuli.

Another interesting aspect of human performance in these experiments is that one of the vowels in each pair is generally perceived to be dominant [8]. The dominant vowel is perceived first and accurately and it is primarily the recognition of the non-dominant vowel which is improved by increases in duration, and pitch differences [8,13]. Although people can recognise individually presented harmonic and inharmonic vowels approximately equally well, the harmonicity of the dominant vowel has been shown to exert a clear influence on the recognition of the non-dominant vowel [3]. It is far easier to recognise the non-dominant vowel when the dominant vowel is harmonically organised than when it is inharmonic. However, the harmonic organisation of the non-dominant vowel has little effect. In other experiments investigating the perception of vowels obscured by broad band masking signals it has been argued that the masked sound may exert its presence by disrupting the patterns of organisation, e.g. periodicity patterns, in some frequency channels [3,13]. One possible conclusion to be drawn from this work is that since the organisation of the masker is crucial in the perception of the masked sound, this indicates that the auditory system may act by suppressing dominant sounds in order to expose masked ones and therefore the effectiveness of this suppression is dependent on the organisation of the dominant sound [13]. These results clearly support a model of auditory processing which achieves the enhancement of target signals largely through the suppression of interfering ones [3].

In this paper it we propose a new model of double vowel perception which is based upon the findings summarised above and upon the physiology of the auditory system, and consider the implications of the model within the wider context of auditory perception. The importance of feedback inhibition in modifying the way incoming signals are processed was a key feature of a model of auditory streaming which we developed previously [5]. Here we explore further the role of top-down inhibition. An important feature missing from previous models of double vowel perception was an account of possible interactions between the recognition and segregation processes. In this model we show how such how such top down influences could help to segregate the sound space more effectively.

2. The Model

Although it is well known that people can distinguish, well above chance, two overlapping vowels, starting and stopping at exactly the same time and based on the same fundamental frequency, the cues which are usually taken to trigger grouping in auditory perception, such as common onset or common fundamental, are clearly of no use within this task. How then might the segregation of such acoustic input occur ? As outlined above, one mechanism which would allow the perception of the dominant and non-dominant vowels in brief double vowel stimuli would be the rapid inhibition of incoming activity associated with the dominant vowel once it had been recognised. The connectivity of the thalamocortical system seems well suited to such processing. The thalamus forms the sensory gateway to the cortex and has a clear role in gating sensory transmission to cortex, in dynamically modifying the receptive fields of cortical neurons and in selectively enhancing parts of the input space as a result of attention [4]. It also seems likely that the recognition of species specific sounds is mediated by thalamocortical processes. In this model we aim to explore the feasibility of the hypothesis that a process of short term suppression operating on the thalamic relay cells might support double vowel segregation and perception, using computational models to show how the thalamocortical auditory system could function in this way.

Although the thalamus has long been thought to have a role in flexibly suppressing or enhancing parts of the incoming signal, it is not yet clear precisely how this is achieved. The model is based on the structure of the thalamocortical auditory system outlined here. The MGB_v is the principal thalamic nucleus of the auditory tonotopic pathway. It is characterised by a tonotopically organised layered structure and is reciprocally connected to the primary auditory cortex, with topographically organised projections [16]. The principal excitatory cells in the thalamus, also known as relay cells, are pyramidal neurons. They receive excitatory inputs from the inferior colliculus and project to the primary auditory cortex thereby forming part of the relay pathway from periphery to cortex. Relay cells also receive excitatory corticofugal projections which synapse on their distal dendrites. In addition they have two principal sources of inhibition; local interneurons and neurons of the thalamic reticular nucleus. Neurons

of the thalamic reticular nucleus are all inhibitory and have extensive lateral interconnections.

The model, shown in figure 2, contains three arrays of neurons which are used to to model activity in the thalamic relay cells (R), thalamic reticular nucleus cells (TRN) and primary auditory cortex (AI) across the tonotopic axis. Connectivity is primarily topographically organised, although we include a small degree of fan out on the projections and do not assume a precise one to one connectivity. There is also wide spread lateral inhibition between TRN neurons, as found in the biological system. It should be noted that since AI feedback is excitatory, the effect on relay cells can be both excitatory and inhibitory, via the TRN. However, the direct AI to relay projections activate metabotropic receptors and generate very slow EPSPs within the relay cells [6]; therefore at the durations with which we are concerned here, this input probably has little importance.

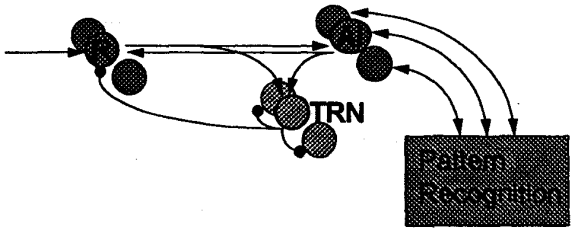

Figure 2. Overview of the model used to explore double vowel recognition.

In order to simulate the behaviour of the system we use an adapted version of the dynamic neuron model described in [7,9]. The time course of synaptic activity is modelled as a low pass filter, as described in [14].

We do not model auditory peripheral processing here, instead we use simplified real valued inputs which resemble the time course and tonotopically distributed patterns of activity found within the auditory system [5]. The stimuli are formed by superimposing two 'vowel' patterns. The vowels each consist of 6 harmonics of the fundamental frequency (100 Hz), centred on the first 3 formants [1]. Although no pitch information is used here, it should be noted that all vowels are based on the same fundamental frequency.

In the absence of any top-down influence, lateral inhibition within the TRN would act to selectively suppress peaks in the activity in the thalamus. If the two vowels were presented at different intensities, then this would be enough to selectively suppress the formants of the dominant vowel. However, when the activity associated with one vowel is not consistently greater than that of the other, then some sort of top-down control of the suppressive process is required. Within the model we include a set of recognisers which selectively respond to each of the vowels. This is consistent with the finding in AI of cells which are responsive to species specific sounds and are affected by learning and prior experience [15]. The vowel template associated with the most active recogniser is used to modulate the AI projections to TRN.

3. A Simplified Double Vowel Experiment

We start with the basic assumption that in order to be consciously perceived, signals must at least activate the primary auditory cortex (AI). A measure of pattern recognition is obtained on the basis of a simple inner product between AI activity and the set of vowel templates receiving the output from AI. A competitive interaction between the recognisers determines which pattern should modulate the AI output activity, thereby effectively determining which of the relay cells should be suppressed. As can be seen in figure 3, the model is capable of substantially reproducing the judgement of human listeners [8].

An important feature of the model is the interaction between the recognition process and the modulation of thalamic response to input stimuli by the inhibition of the relay cells via the TRN projections. When there is no noise, then the effect of such feedback inhibition is not very significant. but when there is background noise, such as in Assman and Summerfield's FS9 vowels [1], then the model has been found to be far more robust with than without feedback inhibition.

4. Discussion

The results we have presented above show that the inhibitory feedback projections from the thalamic reticular nucleus to the thalamic relay cells could have the effect of exposing obscured sounds by suppressing the dominant sound. The suppression process may be partly automatic, but is also likely to be influenced by prior learning, as demonstrated by the use of recognition to gate AI feedback activity. These results are in line with those of [3,13], where it is argued that the organisation of the non-target sound crucially influences our ability to perceive the target sound.

A basic assumption underlying the model is that there is no need to continue to process sensory signals in the same way once they have been received by the cortex, and other potentially useful information may be extracted from the signals if they are processed differently. The TRN seems to be in a position to apply

an inhibitory control over the way in which signals are processed in the thalamus. In turn the TRN is controlled from a number of sources, including the adjacent thalamus and sensory cortex, the reticular formation in response to arousal, and the prefrontal cortex in relation to attentional processes [4]. The TRN therefore seems to provide a way in which a number of areas of the central nervous systems, acting in a competitive or cooperative fashion, can influence the processing of incoming sensory signals.

In the model we have shown that one source of top-down control which may be particularly appropriate in vowel perception, could be some sort of recognition process which ensures that the parts of the input which are associated with a vowel that has already been recognised are suppressed. Although we have not included pitch processing, it is easy to see how the model could be extended to include pitch cues. Pitch appears to be calculated subcortically separately within each frequency channel. The question is how are the components combined to form the pitch percept of a complex sound. If this were to happen cortically, in a way similar to the summary autocorrelation suggested in [10], then the most salient pitch could function in the same way as the most salient vowel, and projections from AI to TRN could ensure that those frequency channels associated with that pitch could be selectively suppressed, thereby exposing a sound with a less salient pitch. Since a clear pitch percept appears to take at least 40 ms to form [12], this would also account for the effects of differences in fundamental frequency (ΔF_0) only being evident at longer durations. This is consistent the comments of McKeown and Patterson who noted that although the ΔF_0 advantage did not hold at short stimulus durations, this did not necessarily imply that pitch was not useful at longer durations [8]. We would expect that other cues, such as direction, frequency modulation and amplitude modulation, may be used in a similar way.

It would appear in the auditory system, as suggested in the visual system [11], that features which emerge as a result of the initial grouping process, may then be used to influence subsequent grouping, i.e. once a clear vowel or pitch begins to emerge for a group of spectral components, then that vowel or pitch can be used to further strengthen the grouping process. This view of auditory scene analysis suggests that the attributes of sounds used to determine grouping may vary with time; that although the primary initial grouping cues are frequency and temporal proximity, once common features begin to emerge from this initial grouping, they too can be used to promote and strengthen the coherence of the grouping. There is therefore some overlap between the sets of factors which cause grouping and those which arise from grouping.

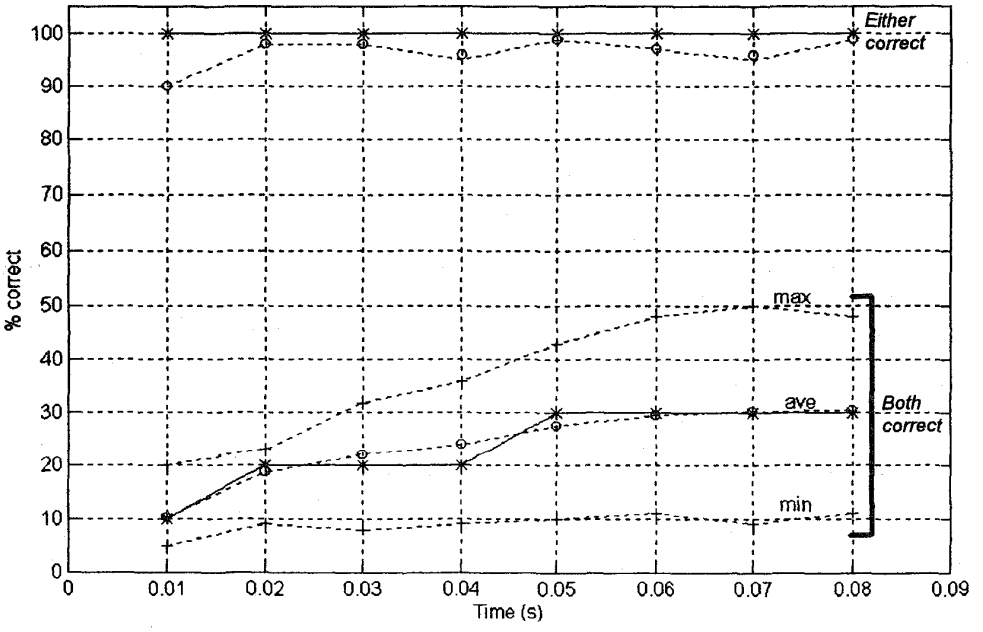

Figure 3 : The probability of correctly identifying both vowels or either vowel in a pair, showing the effect of duration on perceptual accuracy. Model (*, solid line) and human (o, dotted lines) judgements of pairs drawn from 5 vowels EE, AH, OO, OR, ER. Human results are adapted from [8].

Acknowledgements

We would like to thank Ray Meddis for the very interesting discussions which helped to stimulate this work.

References

[1] P.F. Assman, Q.A. Summerfield, "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency". *J. Acoust. Soc. Am.* Vol. 85 Num. 1, 1989, pp327-338.

[2] P.F. Assman, Q.A. Summerfield, "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies". *J. Acoust. Soc. Am.* Vol. 88 Num. 2, 1990, pp680-697.

[3] A. de Chevigne, A. McAdams, J. Laroche,M. Rosenberg, "Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement". *J. Acoust. Soc. Am.* Vol. 970 Num. 6, 1995, pp3736-3748.

[4] E.G. Jones, *Thalamus*, Plenum Press, 1985.

[5] S.L. McCabe, M.J. Denham, "A model of auditory streaming". *J. Acoust. Soc. Am.* Vol. 101 Num. 3, 1997, pp1611-1621. .

[6] D.A. McCormick, M Krosigk, "Corticothalamic activation modulates thalamic firing through glutamate metabotropic receptors". *Proc. Natl. Acad. Sci. USA* Vol. 89, 1992., pp2774-2778.

[7] R.J. McGregor, *Neural and Brain Modelling.* springer-Verlag, 1989.

[8] J.D. McKeown, R.D. Patterson, "The time course of auditory segregation: Concurrent vowels that vary in duration". *J. Acoust. Soc. Am.* Vol. 98 Num. 4, 1995, pp1867-1877.

[9] R. Meddis, M.J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental freqeuncies". *J. Acoust. Soc. Am.* Vol. 91 Num. 1, 1992, pp233-245.

[10] R. Meddis, M.J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification" *J. Acoust. Soc. Am.* Vol. 89 Num. 6, 1991, pp2866-2882.

[11] D. Mumford, "Neuronal architectures for pattern-theoretic problems". *Large Scale Neuronal Theories of the Brain,* Koch, C. , Davis, J. *(ed.s)*, 1994, pp125-152. Bradford Books, MIT Press.

[12] K. Robinson, r.d. Patterson, "The stimulus duration required to identify vowels, their octave, and their pitch". *J. Acoust. Soc. Am.* Vol. 98 Num. 4, 1995, pp1858-1865.

[13] Q.A. Summerfield, J.F. Culling, P.F. Assman, "The perception of speech under adverse conditions: contributions of spectro-temporal peaks, periodicity, and interaural timing to perceptual robustness". *Keele ESCA Conference Proceedings,* 1996.

[14] M.V. Tsodyks, T.J. Sejnowski, "Rapid switching in balanced cortical network models". *Network: Computation in Neural systems* Vol. 6, 1995, pp111-124.

[15] N.M. Weinberger, D.M. Diamond, "Dynamic modulation of the auditory system by associative learning". *Auditory Function,* Edelman, G.M., Gall, W.E., Cowan, W.M. (ed.s). John Wiley and sons. 1988.

[16] J.A. Winer, "The functional architecture of the medial geniculate body and the primary auditory cortex". *The Mammalian Auditory Pathway: Neuroanatomy,* 1992, pp222-409. Springer-Verlag.