MODELLING THE PERCEPTION OF SIMULTANEOUS SEMI-VOWELS

G.F. Meyer[1] and W.A. Ainsworth[2]
Human and Machine Perception Research Centre
Dept of Computer Science[1] and Dept of Commmunication and Neuroscience[2]
Keele University
Keele, Staffs., ST5 5BG, UK
Tel ++44 1782 584111, Fax ++44 1782 713082, email {georg|bill}@cs.keele.ac.uk

Abstract

A model that is able to predict human performance in a simultaneous glide recognition task is described. The model combines a primitive, F_0 guided, segregation stage and a schema driven stage with a heuristic that models whether listeners perceive a single or two simultaneous sounds.

Introduction

Previous studies [1,2,3] suggest that human listeners use simple cues, such as signal harmonicity, speaker location or segmental onset and offset to aid in the segregation of simultaneous sounds. These cues are called 'primitive' grouping cues because they can be applied without prior knowledge. The only heuristic is that segments in an 'auditory scene' that share the same features are likely to be produced by the same speaker. In addition to the primitive segregation process human listeners use high-level knowledge, schemata, to deal with mixtures of sounds [1].

One of the most intensively studied primitive grouping cues is harmonicity. Figure 1 shows human performance for a recognition task involving simultaneous vowels. Each of the panels shows the percentage of pairs that listeners correctly recognise. The stimuli were pairs of the French long vowels /ɑ,e,i,o,u,y/. One of the vowels always had a fundamental frequency (F_0) of 100Hz, the fundamental frequency of the second vowel is plotted along the x-axis. The only primitive segregation cue is the vowel fundamental frequency.

The three panels show subject performance for vowels of 200ms, 100ms and 50ms duration. For signals of at least 100ms duration subject performance improves significantly as the frequency difference between the vowels increases. If signals are only 50ms long no improvement in performance is seen.

The perceptual data is surprising considering the dynamic nature of speech sounds where stationary segments of more than 100ms duration are very rare.

Another important feature that emerges from the data is that humans are able to recognise both constituents of a pair in around 65% of all cases independent of the signal duration and without any grouping cues.

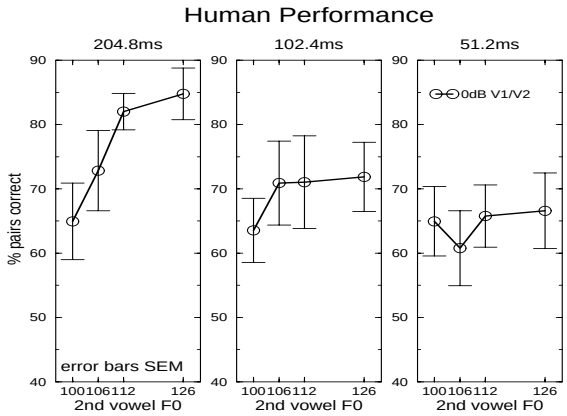

Figure 1; Human recognition performance for pairs of simultaneous vowels. The data is plotted for 204.8ms (left), 102.4ms (centre) and 51.2ms (right) signal duration.

Two key questions arise from these findings

1) how can the reduced human performance for short vowels be reconciled with dynamic speech features?
2) How do human listeners combine the low level grouping cues with high level, schema based, pattern matching strategies.

Glides are an appealing stimulus because segment durations are relatively short and the formant transition introduces dynamicity not seen in steady state vowels. The stimuli are nevertheless relatively easy to create and manipulate.

Human Performance Data

A computer model that is able to replicate human performance on a 'double glide' recognition task is presented. The model is evaluated against experimental data described in detail in a companion paper [4]. A brief summary of the results is given below.

Subjects were presented with simultaneous synthetic glides /jɑ/, /ji/, /wɑ/and /wi/. The relative amplitude of the two glides forming a pair was varied as ratios of 9:1 to 1:9 (-20dB to +20dB). In some of the experiments the fundamental frequencies of both components were set to

100Hz, in others one glide was presented at 100Hz, the other at 150Hz.

This paper is primarily concerned with an evaluation and extension of an exiting F_0 guided segregation model [3,5,6]. The modelling studies consequently focus on data obtained for stimuli with different F_0's.

Subjects were always presented with glide pairs, but were not forced to report the pair heard. In the majority of trials only a single glide is reported to be heard. The proposed model includes this decision process.

The segregation model contains two components. A primitive segregation stage based on the F_0 grouping cue, and a schema driven stage that uses knowledge about the templates rather than primitive segregation cues to segregate speech. The following sections discuss the models in turn.

Segregation with Modulation Maps

Amplitude Modulation maps (AM map) are a physiologically inspired signal representation [7] that allows the segregation of simultaneous voiced sounds. The model is able to replicate human performance data accurately when 200ms windows of speech are processed [2,5].

The processing steps involved are as follows:

1. The speech signal is passed through as 32 channel auditory filter-bank. The model includes a hearing threshold. Filter bandwidths and spacing are matched to perceptual data.
2. The output of each filter is 'demodulated' by half-wave rectification and low-pass filtering.
3. Consecutive 204.8ms frames of the signal are Fourier transformed and the representation shown in figure 3 is obtained.

The AM map representation is two-dimensional with the spectral envelope as the y-axis and the modulation frequencies seen in each channel ac the x-axis. For voiced sounds a characteristic pattern of ridges is seen.

Spectra are recovered from the modulation maps by sampling the ridges at the initial five multiples of the signal pitch. The pitch estimates are supplied and match the signal fundamental.

Model Evaluation

AM maps are used to segregate voiced speech sounds using harmonicity information. Spectra were recovered from overlapping 204.8ms analysis windows, computed every 50ms.

The pattern matching stage is based on the cross-covariance between the extracted spectra and a set of pitch templates obtained by running the model with isolated glides. The stimuli have a fixed segmental structure so that a single value is computed over the full signal duration. Whichever template results in the highest cross-covariance measure identifies the signal.

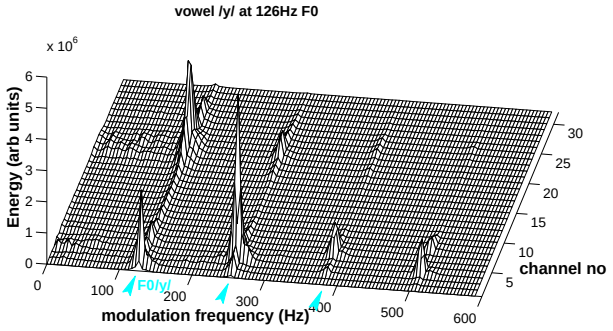

Figure 3: AM map for a single vowel /y/ at 126Hz F_0. The map shows energy as a function of modulation frequency(y-axis) and channel number (x-axis). Channel centre frequencies range from 100Hz to 4.7kHz. Voiced speech sounds form characteristic ridges at multiples of the F_0 of the signal.

In spite of the considerable smearing due to the long window duration, the model recognises 100% of the glide pairs at relative levels up to ±7dB. At ±20dB 60% of pairs are recognised. The model performs significantly better than human listeners on this task, but the matches returned for the (wrong) next best guess tend to be very close to the correct match. The reason for the close proximity of the best two candidate templates is that the AMap response is dominated by the vowel component, which is shared among two candidates (/ji/ - /wi/ and /jɑ/ - /wɑ/).

Schema Based Recognition Model

Human listeners are able to segregate simultaneous sounds, even if no primitive grouping cues are present. Zwicker [8] proposed a 'multiple looks' model, where in a first pass one of the target sounds is identified. The contribution of this sound to the percept is removed and a second look is taken at the remainder.

The schema model is implemented as a simple subtraction stage, which takes the spectral envelope of the glide pair and subtracts the normalised best matching template from the normalised input pattern to obtain a second input to the pattern matching stage.

There is no evidence suggesting that long analysis windows are required to segregate sounds using schemata. The data reported here is based on successive 50ms analysis windows from the auditory filterbank feeding the AM-map. Shorter windows do not lead to performance improvements.

Schema model evaluation

For this limited task this algorithm is effective and, like the F_0 segregation stage alone, results in very good performance (91% avg.) between -7dB and +7dB and 50% pairs correct at ± 20dB. While the model performs well, the spectra are distorted. The first match is a combination of two signals, the second guess is the result of a crude subtraction, which leads to low cross-covariance measures.

Combining the Evidence

The two recognition models produce two matches each, which have to be combined. The final model has to include some decision mechanism that resolves conflicts between the two extraction stages. Both the primitive segregation stage and the schema model perform very well, which leads to the question why human listeners do not make use of this information for glides. One explanation is that auditory scene analysis does not take place, but in the light of evidence from previous studies which show it to be a very general process, this seems very unlikely [1,2,3,4,8]. A more plausible explanation is that, unless there is clear evidence that a second glide is present, listeners expect a single signal to be present. Both segregation stages have difficulty in supplying this evidence.

Average cross-covariance (r_{xy}) values are shown in fig. 4. The top panel shows r_{xy} values when the vowel components of the two glides differ while the bottom panel shows data for glide pairs that have the same vowel component. The schema model data is based only on the first, not the second match, which is always below all other matches, while both matches from the F_0 segregation stage are used.

Both segregation models produce candidate matches with associated distance values, which are used as the basis of the conflict resolution stage shown schematically in figure 5. Output from the model that produces the best combination of matches is chosen.

Estimating the number of signals present

An important decision for human listeners to make is whether a single glide or a pair is present. This decision can only be based on the similarity of the complex (i.e. the primary schema model output) to a isolated glide. The heuristic chosen is to compare the template match for the primary schema model output, after normalisation, with a set of matches that would be expected if the model was driven with isolated glides. If the set of matches includes no secondary matches that are higher than those expected for isolated signals, the model assumes that only one signal is present.

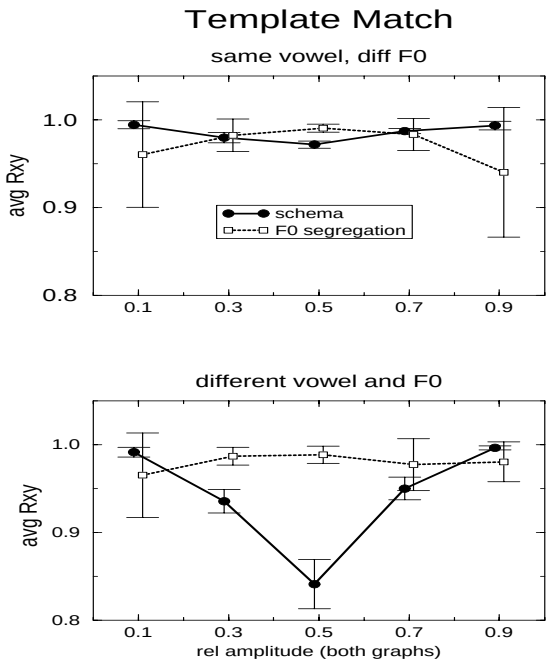

Figure 4: Average distance measures for the two models for two experimental conditions. For details refer to the main text.

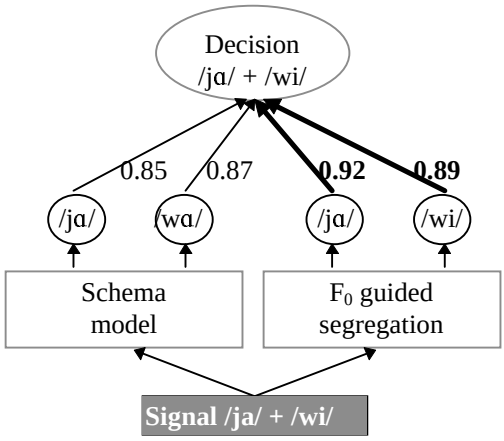

Figure 5: Schematic representation of the decision process. Both models work in parallel and produce candidate matches. For details refer to the main text.

Combined Model Evaluation

Performance predicted by the combined model is shown in figure 6. The performance data is plotted for the two categories of signals discussed previously. The model predicts that human listeners hear isolated glides at the two extremes of the relative amplitude spectrum, but that, when both components are equally loud, both components are heard and identified correctly. The emerging picture is qualitatively similar to human performance but overall correct rates are much higher (100% vs. 60%) and the difference between glide pairs with same

and different vowels, seen in human listeners, is not visible, fig. 6 (cf. fig, 2b,d and 3 in [4] for human data).

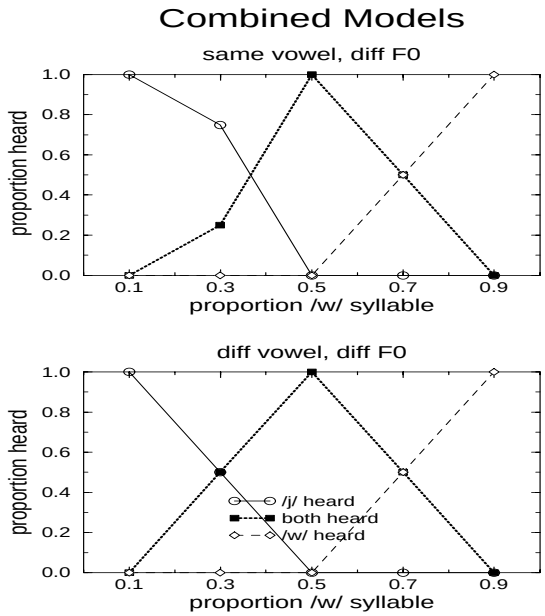

Figure 6: Recognition performance predicted by the deterministic model. Refer to the main text for details.

Adding noise to the system

The model described in the previous section performs significantly better than real listeners. This is to be expected because the signal representation in the auditory system and presumably the decision making processes are noisy. This has little effect if decisions are clear-cut, but the distance measures show that both decision making models are based on candidates which are very close to each other. Noise would increase the likelihood of phoneme confusions and also make the decision whether a single glide or a pair is present less reliable. Figure 7 shows performance data for a model when gaussian noise ($\bar{x}$-0, σ=0.05) is added to both processes. For all stimuli heard as pairs, recognition performance reduces to about 60%, which is in line with human data.

Conclusions

We propose an abstract model that includes three components, a primitive F_0 guided segregation stage, a high level, schema based, segregation stage, and a decision making process that interprets the data from both segregation stages.

Both the low-level segregation stage and the schema model outperform human listeners, but the metrics underlying the decision are very fragile. A non-deterministic model is able to replicate the main effects seen in human data.

The modelling study confirms the assumption that long analysis windows are used in the F_0 guided segregation stage.

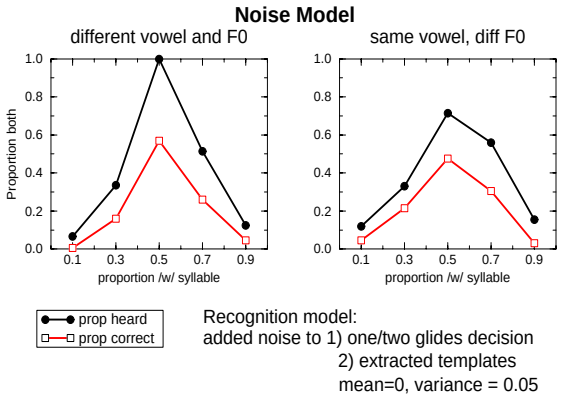

Figure 7: Probablistic segregation model. Recognition performance reduces to around 66% of stimuli perceived as containing two signals. One glide alone is 'heard' more often if both vowel components are the same. The model matches human data qualitatively

Acknowledgements

We are grateful to Frederic Berthommier who contributed greatly to the development of the AMap model. This work is funded by EPSRC grant GR/L05655.

References

[1] A.L. Bregman (1990) "Auditory Scene Analysis" MIT Press, Cambridge MA, 1990

[2] P.F. Assmann and Q. Summerfield (1990) Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies" J Acoust Soc Am 91, 233-245.

[3] F. Berthommier and G.F. Meyer (1995) "Source Separation by a Functional Model of Amplitude Demodulation" Proc Eurospeech '95 135-138.

[4] W.A. Ainsworth and G.F. Meyer (1997) "Preliminary Studies on the Perception of Double Semi-Vowels" Proc Eurospeech '97, this volume.

[5] G.F. Meyer (1996) "Expanded Signal Representations for Auditory Scene Analysis " Proc. IOA 18 3-10.

[6] G.F. Meyer and F. Berthommier (1996) "Vowel Segregation with Amplitude Modulation Maps: A Re-Evaluation of Place and Place-Time Models" Proc ESCA workshop on the Auditory Basis of Speech Perception, 212-215.

[7] C.E. Schreiner and G Langner (1988) "Periodicity Coding in the Inferior Colliculus of the Cat. II Topographical Organization" J Neuripysiol. 60 1832-1840.

[8] U.T. Zwicker (1984) "Auditory recognition of diotic and dichotic vowel pairs". Speech Communication, 3, 365-277.