An Adaptive Sampling Procedure for Speech Perception Experiments

Geoffrey Stewart Morrison

Department of Linguistics
University of Alberta, Edmonton, Alberta, Canada
gsm2@ualberta.ca

Abstract

Synthetic speech perception experiments may make use of several acoustic dimensions in order to adequately model listeners' perception; however, the number of stimuli increases exponentially as dimensions are added. A relatively large number of identification responses per stimulus are needed in the vicinity of category boundaries in order to model the boundaries with reasonable accuracy. Fewer responses per stimulus are needed to model portions of the stimulus space where a single response category predominates. Rather than collecting the same number of responses for each stimulus, an experiment can therefore be shortened via adaptive sampling. An adaptive sampling procedure is described. After an initial pass through the stimuli, the procedure uses a logistic regression model to select stimuli to resample in subsequent rounds. Results of simulations indicated that the number of trials in the experiment could be reduced by a third without substantially affecting the results.

Index Terms: adaptive sampling, speech perception

1. Introduction

A typical speech perception experiment involves creating a set of synthetic speech stimuli whose acoustic properties form a multidimensional matrix, randomly presenting each stimulus a fixed number of times, and, at each presentation, having a listener classify each stimulus as one of a number of speech sound categories. Data consist of the proportion of responses for each category given to each stimulus. A simple experiment might involve a two-dimensional matrix and two speech sound categories, e.g., equally spaced vowel duration steps on one dimension and equally spaced first formant (F1) steps on another dimension, covering the range of F1 and duration values between English /i/ and /ɪ/. More complex experiments may involve a larger number of response options and a larger number of stimulus dimensions. Several acoustic dimensions may be necessary to adequately model listeners' perception, but as the number of dimensions increases, the number of stimuli increases exponentially.

From the perspective of building an accurate unbiassed statistical model of a listener's speech categorisation, it is desirable to obtain a large number of responses for each stimulus. With a larger number of samples, there will be greater resolution in the proportional responses for each category. Unfortunately, collecting a large number of responses from human participants is time consuming, the participants can quickly become fatigued, and they may be reticent to return to participate in subsequent sessions in longitudinal or multiple-condition studies. The present paper describes an adaptive sampling procedure which was developed in order to make more efficient use of participants' time whilst still obtaining a reasonable degree of resolution in the proportional responses. The procedure focuses on boundaries and has some similarities with up-down methods [1]. For a different approach to adaptive sampling focussing on best exemplars see Evans & Iverson [2].

2. Stimulus Set

The adaptive sampling procedure was initially developed for use with an experiment investigating the perception of English /i/, /ɪ/, /e/, /ɛ/, and Spanish /i/, /ei/, /e/ [3]. The procedure will be described using the stimulus set from this study as a concrete example. There were a total of 90 synthetic vowel stimuli covering three acoustic dimensions. The duration dimension had three points [80, 95, 110 ms]; the F1–F2 dimension had ten points, the first and second formants (F2) at the beginning of the vowel covaried forming a diagonal in the F1–F2 space [F1: 283–580 Hz in 33 Hz steps, F2: 2090−1730 in 40 Hz steps]; and the vowel inherent spectral change (VISC [4]) dimension had three points, from the beginning of the vowel to the end F1 and F2 either diverged, remained flat, or converged [ΔF1: −99, 0, +99 Hz, ΔF2: +120, 0, −120 Hz]. The number of stimuli had been winnowed from a larger stimulus space, by combining the F1 and F2 dimensions and reducing the number of points on each dimension; however, the stimuli were embedded in words in carrier sentences and in pilot tests it took listeners approximately half an hour to identify each stimulus four times (360 trials). The goal was to develop a sampling procedure which would give a resolution comparable to six responses per stimulus within the half hour time frame.

3. Adaptive Sampling Procedure

3.1. Basic procedure

The essential principle underlying the procedure is that certain stimuli will not need to be sampled a large number of times because they fall near the middle of a listener's perceptual space for a given category, and will therefore always be identified as that category. For example, if a stimulus is in the middle of the perceptual space for a listener's /i/ category, then the listener will always identify this stimulus as /i/; thus irrespective of the number of responses the listener gives to this stimulus, the proportion of /i/ responses for this stimulus will always be 1. Hence, once portions of the perceptual space which are far from boundaries have been located, there is no need to obtain further responses in those areas. On the other hand, stimuli near category boundaries may be identified as one category on one occasion, and as another category on another occasion. For example, a stimulus may be identified as /i/ two thirds of the time and as /ɪ/ one third of the time, and a neighbouring stimulus may be identified as /i/ half the time and as /ɪ/ half the time. In order to determine the proportion of /i/ responses with reasonable resolution such stimuli must be sampled a considerable number of times.

The procedure consists of the following steps:

1. All the stimuli are sampled twice, i.e., all the stimuli are presented in two blocks (once in each block) and the listener gives a identification response on each trial (180 responses).
2. A logistic regression model is fitted to the response data, and the predicted probabilities for each category are calculated for each stimulus.
3. The error between the predicted probability and observed proportion for each category for each stimulus is calculated.
4. Half of the stimuli, primarily those with the largest error scores, are resampled (45 responses, see Section 3.2).
5. Steps 2 through 4 are repeated three more times.

This procedure results in 360 trials, and each stimulus is sampled a minimum of twice and a maximum of six times. After two rounds, a stimulus which receives two /i/ responses and is surrounded by stimuli which receive two /i/ responses is unlikely to be near a category boundary. This stimulus will have an observed proportion of /i/ responses of 1, and a predicted probability for /i/ close to 1. This stimulus will therefore have a low error score, and is unlikely to be resampled in subsequent rounds. In contrast, a stimulus which receives two /i/ responses but is adjacent to stimuli which receive /ɪ/ responses, will have an observed proportion of /i/ responses of 1, but will have a predicted probability for /i/ that is somewhat less than 1. This stimulus will therefore have a higher error score, and is more likely to be resampled in subsequent rounds. A stimulus which receives one /i/ response and one /ɪ/ response could have a small error between observed and predicted values, but, especially in a multidimensional stimulus space and with multinomial response categories, it is more likely to have a relatively large error. In practice, the vast majority of stimuli near category boundaries receive relatively high error scores, and stimuli far from category boundaries receive low error scores.

An alternative procedure which resampled the stimuli with predicted probabilities furthest from 0 and 1 was also explored. Selecting stimuli using this criterion gave similar results to using the highest-error-score criterion, but the latter offered the advantage of a stronger mistake amelioration feature: A mistake where a listener accidentally presses the wrong button is likely to increase the error score for the stimulus on which the mistake was made. Using the highest-error-score criterion, that stimulus is therefore more likely to resampled, leading to a reduction in the effect of the mistake.

The multinomial logistic regression algorithm was based on Haberman [5] and its use in speech perception experiments is described in Nearey [6, 7]. The model fitted was a simple first-order model ($V + V\times$F1 $+ V\times\Delta$F1 $+ V\times$dur) containing one bias and three stimulus-tuning coefficients for each vowel category. Stimulus-tuning coefficients consisted of F1-tuning with initial formant values for F1 entered in Hertz (since F2 covaried with F1 it was redundant), ΔF1-tuning, with change in F1 value from the beginning to the end of the vowel entered in Hertz, and duration-tuning, with vowel duration values entered in milliseconds. All stimulus properties were treated as continuous variables. The number of each type of coefficient in the fitted model was actually one less than the number of categories, the coefficients for the last category being redundant and calculable as minus the sum of the coefficients for the other categories. A simple model is preferred to avoid overfitting the sparse data sets, especially near the beginning of the adaptive sampling procedure. An overfitted model may wrap around fluctuations in the data sets due to course sampling and give lower error scores to stimuli near boundaries than would the optimal model. In simulations, use of a quadratic model resulted in unstable results with high variances for the coefficients in the final model. Using an underfitted model during adaptive sampling will be less efficient than the optimal model, but will not obliterate more complex patterns in the data which may be captured by fitting a more complex model to the final results. If the model makes a linear approximation of a curved boundary then some stimuli will be a poor fit to the model because the model is underfitted; however, this will lead to these stimuli being resampled and the curved boundary will still be represented in the final data set.

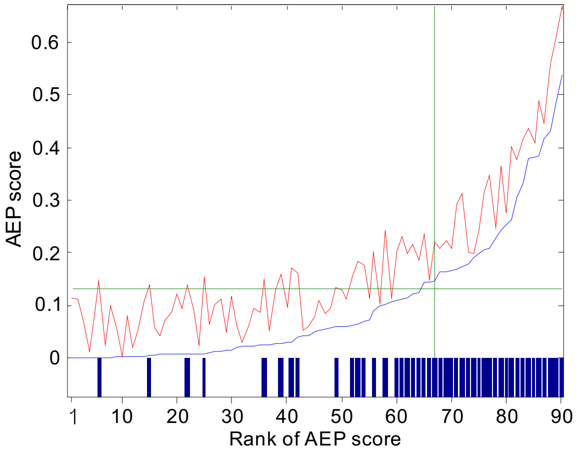

Figure 1 *Example of selection of stimuli to be resampled on the basis of absolute errors in proportions (AEP) for a model fitted to two responses per stimulus.*

3.2. Selecting stimuli to resample

Rather than simply resampling the 45 stimuli with the highest error scores, the stimuli to resample were chosen such that those with higher error scores were most likely to be resampled but those with lower error scores also had some probability of being resampled. This ensured that listeners heard some reasonably good examples of the vowel categories in each round. Good examples provide the listeners with anchors against which to compare more ambiguous stimuli, good examples will also be easy to identify and thus be reassuring for the listeners. The stimuli to resample were selected stochastically as follows:

1. The stimuli were ranked in ascending order of their error scores, resulting in a sequence which increased in an approximately exponential manner (see Figure 1).
2. The error score of the 67th stimulus of the 90 ranked stimuli was obtained. (Vertical line in Figure 1)
3. Integers from 1 to 90 were randomly permuted then divided by 90 and multiplied by the error score of the 67th ranked stimulus. This generated a sequence of random numbers with the highest number being equal to the error score of the 67th ranked stimulus.
4. The sequence of ranked error scores and the sequence of random numbers were added. (Noisy line in Figure 1)
5. Stimuli with error-plus-random scores of greater than the median value were selected for resampling. (The median value is represented by the horizontal line in Figure 1. The

stimuli selected for resampling are indicated by the bars at the bottom of the figure.)

Half the stimuli are resampled. All the stimuli have a non-zero probability of being resampled which increases with their error score, and the quarter of the stimuli with the worst fit are guaranteed to be resampled.

3.3. Error measures

Standard error measures such as *Root Mean Squared* (RMS) error are usually calculated assuming that each stimulus is sampled an equal number of times, which is not the case for the adaptive sampling procedure. Ad hoc error measures used instead were the *Absolute Errors in Proportions* (AEP) for individual stimuli, and the *Sum of the Absolute Errors in Proportions* (SAEP) for the stimulus set.

The AEP for a stimulus is calculated as half the sum of the absolute difference between the observed proportion of responses and the predicted proportion of responses for each category for that stimuli, or equivalently as half the sum of the absolute difference between the observed and predicted number of responses for each category divided by the total number of responses for that stimulus:

$$AEP_{stim} = \frac{\sum_{cat} \left| NumObserved_{cat} - NumPredicted_{cat} \right|}{2 \times NumResponses_{stim}}$$

The theoretical minimum and maximum values for AEP are 0 and 1 (the scaling factor of ½ was introduced to make the maximum value 1). An AEP value of 0 indicates a perfect fit between the observed responses and the model's fitted responses, and an AEP value of 1 indicates a complete mismatch (e.g., if the participant always responded with one category, and the model predicted a probability of zero for that category). The SAEP for the stimulus set is calculated as the sum of the AEP for all stimuli.

An alternative error measure could have been to calculate errors of fit on the basis of differences between observed and predicted logit values. The error measure based on proportions was preferred since errors which would be the same size in logistic values, are smaller in proportion values when they are close to proportions of 0 and 1 relative to when they are near proportions of .5, and this weighting was advantageous because the error measures were being used as a criterion to select stimuli that were near category boundaries.

4. Simulations

To obtain test data, the full set of stimuli were presented in random order in six blocks (540 trials), and on each trial the stimulus presented was identified by a single listener as one of the four English vowel responses. A first-order logistic regression model was fitted to the whole data set (a territorial map based on this model is given in Figure 2). The a posteriori probabilities from this model were used as population parameters in a multinomial sample generator which generated 100 simulated response sets of six responses per stimulus. Simulated responses were generated independently for each stimulus. To generate a single simulated response for a stimulus, the sample generator chose one of the four English vowels /i/, /ɪ/, /e/, /ɛ/, the probability of choosing a particular response category on each occasion being dependent on its a posteriori probability for that stimulus.

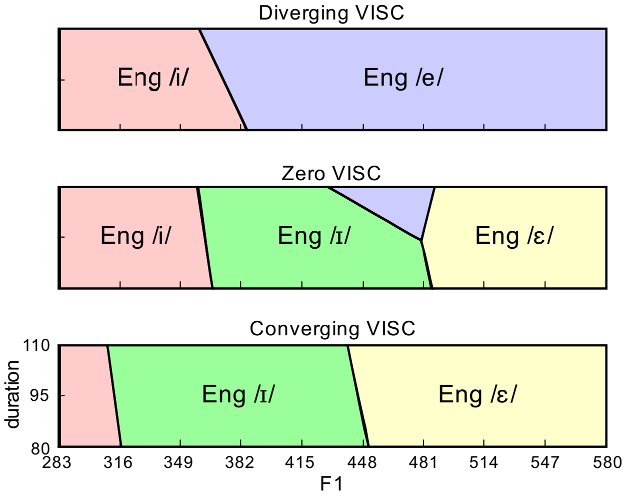

Figure 2 *Territorial map based on logistic regression model fitted to original test data.*

Whole-set logistic regression models were fitted to each of the 100 simulated response sets and the SAEP and coefficient values saved. Models based on the final set of responses selected by the adaptive sampling procedure were fitted to the same 100 simulated response sets. The first two simulated responses to each stimulus were used in both models, but subsequent simulated responses for a stimulus were only used in the adaptive model if that stimulus was selected for resampling. The whole-set models were compared with the adaptive models: for each sample set the difference between the logistic regression coefficient values for the whole-set model and the adaptive model were calculated, and these were used as the test statistic in paired-sample *t*-tests.

Different variants of the adaptive procedure were tested using different criteria for selecting the stimuli to resample and different levels of complexity for the logistic regression model. The version of the adaptive procedure described above was selected as giving the closest results to the whole-set model. Numerical comparisons between the whole-set model and this version of the adaptive model are presented below.

Table 1 presents the results of comparisons between the whole-set and the adaptive model for the simulated response sets. The difference in SAEP between the models was not significant. The differences between models for several coefficients were significant; however, the size of the difference was small, none of the mean differences had magnitude greater than 4.5%. Three of the four significant differences were related to a single response category, /i/, and were therefore not independent of each other: The magnitudes of the bias and the stimulus-tuned coefficients for /i/ all decreased by similar amounts (3.3–4.4%) indicating a slight reduction in the estimate of the rate at which responses changed from /i/ to other categories, but little change in the location of the boundary (if the size or direction of the change in the bias had differed from the size of the change in the stimulus-tuned coefficients, then the modelled location of the boundary would have changed).

In order to test the sampling method on a wider set of simulated data that might reflect a wider range of listeners, the data set was perturbed in several ways. The coefficient values from the logistic regression model based on the original data were reduced to 25% of their original values, and used to

Table 1. *Comparison of error scores and coefficient values across sampling procedures*

Error or Coefficient	Sampling Procedure						
	Whole-Set	Adaptive	Difference				
	Mean	Mean	Mean (sd)		%	*t*(99)	*p*
SAEP	6.813	6.793	−0.020	(0.426)	−0.3	−0.471	.6386
i	34.113	32.923	−1.190	(1.349)	−3.5	−8.823	.0000 **
ɪ	7.141	7.074	−0.068	(0.906)	−0.9	−0.748	.4562
e	−8.147	−8.181	−0.034	(0.792)	+0.4	−0.426	.6708
i×F1	−0.077	−0.074	0.003	(0.003)	−3.6	10.047	.0000 **
ɪ×F1	−0.007	−0.007	0.000	(0.002)	−2.0	0.850	.3975
e×F1	0.012	0.012	0.000	(0.001)	−1.8	−1.742	.0846
i×ΔF1	−2.028	−1.939	0.089	(0.164)	−4.4	5.461	.0000 **
ɪ×ΔF1	1.510	1.486	−0.025	(0.137)	−1.6	−1.802	.0746
e×ΔF1	−3.445	−3.384	0.061	(0.189)	−1.8	3.217	.0018 **
i×dur	−0.037	−0.036	0.001	(0.006)	−3.3	2.020	.0461 *
ɪ×dur	−0.020	−0.021	−0.001	(0.004)	+3.3	−1.612	.1101
e×dur	0.044	0.045	0.001	(0.005)	+1.2	0.999	.3202

* significant at α = .05, ** significant at α = .0038 equal to .05 after a Bonferroni correction for 13 tests
% Percentage differences indicate differences in magnitude which are towards zero if negative and away from zero if positive

generate a further series of 100 sample sets. SAEP was significantly higher for the adaptive compared to the whole-set models [mean 20.425 vs 18.857, *t*(99) = 18.823, *p* < .0038], but none of the coefficient values had significant differences. Another series of 100 sample sets was generated on the basis of the original model, but 25% of the responses were replaced by responses generated at random with each response category having an equal probability irrespective of stimulus properties. SAEP was significantly higher for the adaptive compared to the whole-set models [mean 25.426 vs 24.114, *t*(99) = 5.438, *p* < .0038]. The mean difference in *i*, and *i*×F1 coefficient values between the adaptive and the whole-set models were also significantly different [*i* mean 7.612 vs 7.129, *t*(99) = 5.208, *p* < .0038; *i*×F1 mean −0.016 vs −0.017, *t*(99) = 5.208, *p* < .0038], the magnitude of both these differences was 6.8%.

5. Conclusion

On the basis of the simulations, it was decided that any small differences in the accuracy of results were immaterial compared to the benefits accrued by presenting the participants with a shorter experiment, 360 trials rather than 540. The adaptive sampling procedure as described above was therefore adopted for use in data collection in the study of the perception of English /i/, /ɪ/, /e/, /ɛ/, and Spanish /i/, /ei/, /e/. Individual participants took between 20 and 40 minutes to complete the perception experiment, and participant retention was very high: of the 95 participants who were asked to participate in two or more experiment sessions (e.g., one experiment giving English responses and one experiment giving Spanish responses to the same stimuli), only 3 dropped out after the first session.

The best fitting logistic regression model for the data in the study was not restricted to the linear model used in the selection of stimuli to resample: For two groups of first-language Spanish listeners (a monolingual and a bilingual group), the best fitting model for their Spanish vowel category responses had ΔF1 coded as three discrete levels. This allowed for a non-linear relationship between VISC and the model's predictions for Spanish vowel category responses.

It may be possible to increase the relative reduction in the number of stimuli sampled, particularly in experiments with the same number of dimensions, but with a higher stimulus density or greater maximum number of samples per stimulus. An additional reduction could be achieved by running the initial pass on a low density matrix, then switching to a high density matrix using furthest from 0 or 1 stimulus selection.

6. Acknowledgements

Thanks to Bronwen G. Evans, Paul Iverson, Terrance M. Nearey, and two anonymous reviewers for comments on earlier versions of this paper (any errors or shortcomings remain the responsibility of the author). This work was supported by the Social Sciences and Humanities Research Council of Canada.

7. References

[1] Levitt, H. "Transformed up-down methods in psychoacoustics", J. Acoust. Soc. Amer., Vol. 49, 1971, p 467–477.

[2] Evans, B. G., and Iverson, P. "Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences", J. Acoust. Soc. Amer., Vol. 115, 2004, p 352–361.

[3] Morrison, G. S., L1 & L2 production and perception of English and Spanish vowels: A statistical modelling approach. Doctoral dissertation, University of Alberta, 2006.

[4] Nearey, T. M., and Assmann, P. F. "Modeling the role of vowel inherent spectral change in vowel identification", J. Acoust. Soc. Amer., Vol. 80, 1986, p 1297–1308.

[5] Haberman, S. J., Analysis of Qualitative Data. Vol. 2, Academic Press, New York, 1979.

[6] Nearey, T. M. "The segment as a unit of speech perception.", J. Phonetics, Vol. 18, 1990, p 347–373.

[7] Nearey, T. M. "Speech perception as pattern recognition", J. Acoust. Soc. Amer., Vol. 101, 1997, p 3241–3254.