SAME PHONEMIC SEQUENCE, DIFFERENT ACOUSTIC PATTERN AND GRAMMATICAL STATUS. A MODEL

Marco A. Piccolino-Boniforti*, Bogdan Ludusan**, Sarah Hawkins***, Dennis Norris****

*Dept. of Linguistics, University of Cambridge; **Dept. of Physical Sciences, “Federico II” University; ***Medical Research Council Cognition and Brain Sciences Unit, Cambridge

1. map55@cam.ac.uk; *ludusan@na.infn.it; ***sh110@cam.ac.uk; ****dennis.norris@mrc-cbu.cam.ac.uk

1. ABSTRACT

We aim to build a computational model that will help elucidate how humans predict grammatical structure from acoustic-phonetic detail.

The first step is a proof-of-concept model to distinguish between true and pseudo morphological prefixes in English words, such as discolour, in which dis is a true prefix, and discover, in which dis is a pseudo-prefix. Both words have the same first four phonemes, /dɪsk/ but linguistic and phonetic analyses show that pronunciations of pseudo prefixes tend to have a weaker rhythmic beat than pronunciations of true prefixes have (Ogden et al. 2000; Baker 2008; Baker et al. 2007a) and that these differences affect intelligibility of sentences in noise (Baker 2008; Baker et al. 2007b).

The present work uses Baker’s original speech corpus and aims to simulate aspects of her observed results. The computational model comprises two main parts. The acoustic signal is first processed within a cochlear model (Patterson et al. 1988; Meddis 1986) that introduces non-linearities in frequency and loudness. The cochlear output is then transformed into an auditory primal sketch (APS, Todd 1994) which simulates perception of amplitude modulation at various temporal resolutions within the auditory system. This representation identifies successive acoustic events in the signal and their so-called relative prominence, a measure that combines amplitude and duration. In the second stage of the present model, the output of the auditory primal sketch is input to a classifier (target class: true vs. pseudo morpheme). Two classifiers are compared, the popular support vector machine (SVM, Vapnik 1995), and the relevance vector machine (RVM, Tipping 2001). The latter seems to display more interesting properties for the simulation of cognitive processes.

The present work reports simulations that compared: 1) RVM vs. SVM; 2) APS based vs. energy based vectors; 3) cochlear-model based versus non-cochlear-model based APS vectors. Model performance was measured both in terms of classification accuracy and model sparsity.

Results show that both RVM and SVM assign the data to the correct true vs pseudo morphological category at well above chance. According to a mixed-effects ANOVA (main factor: RVM vs. SVM; random factor: subject) accuracy difference is just marginally significant. However, the RVM obtains a much sparser representation than the SVM. Comparing APS vs signal energy accuracy using an RVM classifier, energy performs better. Sparsity is not different. All other parameters being equal, and using an RVM classifier, the cochlear model improves the accuracy of the APS compared to the non-cochlear model version. The cochlear model also achieves greater model sparsity.

These results suggest that true prefixes can be reliably distinguished from pseudo prefixes based on the systematic differences in their acoustic patterns, confirming Baker et al.’s (2007a) findings. Both RVM and the cochlear model show clear advantages in terms of accuracy and/or sparsity. The poorer performance of the auditory primal sketch seems to be linked to the kind of vectors adopted, each one containing a variable number of events.
2. INTRODUCTION

Natural speech contains significant acoustic-phonetic detail that systematically distinguishes linguistic structure and functions (e.g. words, grammatical structure, and particular communicative functions) but that is not reflected in a standard phonemic transcription of the speech (Local 2003). The fact that phonemic transcriptions lack phonetic detail often results in ambiguities in phoneme sequences that machine recognizers typically resolve by using a higher-order ‘language’ model—i.e. without further reference to the signal. However, native listeners are sensitive to much of this acoustic-phonetic detail: the knowledge required to interpret the information is indeed higher level, but the information itself lies within the physical signal (see Hawkins 2003 for a review).

The distinction between true and pseudo morphological prefixes in English words provides one example. When the syllable mis or dis begins a word, it may be either a true (productive) prefix, as in mistimes, discolour, distasteful, in which removing the first syllable results in more or less the opposite meaning; or it may be a pseudo prefix, as in mistakes, discover, distinctive, in which removing the first syllable does not produce a word with the opposite meaning (though it might once have done). The same is true for various other prefixes, but mis and dis (and the following consonant) are the focus of this study. These syllables have the same first four phonemes (e.g. /dɪsk/ for discolour and discover) but their fine spectro-temporal structure differs in predictable ways (Baker et al. 2007a). Heard in sentences or phrases in noise, the words are less intelligible when a syllable of the wrong morphological type replaces the original initial syllable, compared with when a syllable of the right type is spliced into the same position (Baker et al. 2007b).

The data reported by Baker et al. (2007a) show both absolute and relative differences in segment durations and some spectral properties of these syllables associated with morphological status. Broadly, the true prefix has a heavier rhythmic beat than the pseudo prefix, presumably because the true prefix is generally longer with a more peripheral (‘clearer’) articulation, and has a higher proportion of periodic to aperiodic energy: [s] takes up a greater proportion of the total [ɪs] duration in the pseudo prefix than in the true prefix. When the fourth phoneme in the sequence is a stop, as in the above examples, its VOT (aspiration after the transient) is longer after a true prefix. (Strictly, this stop is part of the word’s stem, rather than part of the prefix.) In short, the acoustic properties combine to produce complex yet distinct spectro-temporal patterns that might form integrated perceptual units. Such units could function as distinct ‘auditory objects’ signalling morphological status, regardless of phonemic identity. The perceptual advantage would be to restrict word searches to the appropriate monomorphemic or bimorphemic word sets.

We are developing a model architecture which exploits acoustic detail in order to determine the grammatical status of morphemes which are transcribed with the same phonemic sequence. While adopting a Bayesian perspective about the nature of the speech recognition process, in accordance with the principles outlined in Norris & McQueen (2008), it puts particular stress on the importance of low level processing of the speech input by using the auditory primal sketch (Todd 1994), and on the hybrid episodic-abstract nature of higher level representations by using sparse Bayesian learning techniques (Tipping 2001), specifically those which allow actual input vectors to be retained as candidates for prototypes (the relevance vector machine, RVM, Tipping 2000).
3. THE INPUT DATA

The data used for all simulations was collected by Baker (Baker et al. 2007a). It comprises 1000 mis- or dis- prefix tokens with the inclusion of the onset consonant for the following syllable, thus giving a misC- or disC- pattern, from words of the kind exemplified above. The speech is fast conversational, quasi-spontaneous scripted dialogue, well balanced in terms of prefix type (500 misC-, 500 disC-), number of speakers (5 speakers x 200 tokens), stress borne by the word (500 nuclear, 500 post-nuclear), morphological status (500 true, 500 pseudo), and phonetic context (identical except for the critical word). There are 5 tokens of each speaker/word/stress condition.

For all experiments, 80% (800) of the tokens were used for computing training vectors and 20% (200) for computing test vectors. Once split between train and test, the order of input vectors was randomised, but for consistency the same random seed was used for all simulations. In each simulation, 5 speaker-dependent models were built with the training data (160 vectors per model). The models were then tested on the corresponding test data (40 vectors per model). Models based on the whole training set were also built (800 vectors), and tested on the whole test set (200 vectors). The statistical analyses presented here refer to the speaker-dependent models. Since the results of the speaker-independent models could not be assessed by using the same statistical tests, we preferred to omit them to prevent drawing wrong conclusions as to their significance. Qualitatively, however, they are similar to the speaker-dependent ones, both in terms of accuracy and relative sparsity (see Section 5 for the definition of these measures).

4. THE MODEL

4.1 General structure

The model has lower- and higher-level components. The lower-level component can be thought of as transforming the acoustic input signal into an integrated auditory object. This object is then passed to the higher-level component for Bayesian classification as either a true or a pseudo prefix.

4.1.1 The lower-level model

The lower-level model is based on a multi-resolution model of rhythmic grouping, the auditory primal sketch (Todd 1994). This partially parallels edge detection in vision (Marr 1982). In speech, edges correspond to abrupt acoustic events e.g. at the boundaries between different excitation types, or where changes in vocal-tract shape produce large changes in spectral shape or waveform amplitude over a few (10-40) milliseconds. The model is realised as a bank of Gaussian filters of increasing width through which the output of a cochlear model is passed. The cochlear model is based on the basilar membrane model of Patterson et al. (1988) and the hair cell model of Meddis (1986).

When represented in terms of time vs. filter width, the model output is called a rhythmogram, illustrated for the two [mɪst] prefixes of mistimes and mistakes in Figure 1 (see caption for details). The rhythmogram represents the input signal in terms of the relative prominence of each event with respect to the widest filter width considered. Conversely, the narrowest filter width influences the number of events picked up by the model.

A more compact, slightly less informative, representation of the rhythmogram, obtained by integrating over all filter outputs for a single event, was devised by Lee & Todd (2004).
This returns a vector of events, \textit{pr-scores} (prominence scores). Each \textit{pr-score} is represented by a scalar which reflects event prominence relative to other events in the signal.

The auditory primal sketch

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{auditory_primal_sketch.png}
\caption{The auditory primal sketch. \textit{Left:} [mɪst] from “mistakes”. \textit{Right:} [mɪst] from “mistimes”. Within each of these: Lower panel: waveform; middle panel: rhythmogram without cochlear model; upper panel: rhythmogram with cochlear model. Prominence scores (\textit{pr-scores}) are obtained by integrating over contiguous black squares (peaks).}
\end{figure}

The main advantages of the auditory primal sketch are: a) it exploits perceptual integration to create an auditory object that b) has accessible internal structure reflecting significant acoustic events via its multi-resolution representation of prominence, which is c) produced by analysing duration and intensity information in ways that simulate cochlear and cortical processing. Importantly, the multi-resolution properties of the rhythmogram are also suited to capture dependency relations between events (Todd & Brown 1996).

4.1.2 The higher-level Bayesian classification model

Bayesian principles have long been understood by many as fundamental to successful speech recognition, both for machines and humans (see Scharenborg et al. 2005 for an attempt to unify the two research areas). Classification of any given acoustic signal is strongly affected by context (see Hawkins 2004 for a review); a formulation of spoken word recognition as a Bayesian problem accounts elegantly for many effects that in other modelling frameworks would require special treatment (Norris & McQueen 2008).

The higher-level part of the model uses the RVM, which is a particular instance of a sparse Bayesian method (Tipping 2000, 2001).

RVM is a machine learning technique which shares some properties with the popular support vector machine (SVM, Vapnik 1995), but differs in some important respects.

Both SVMs and RVMs classify by storing a subset of the vectors used for training, rather than the full training set or summary statistics of all the observed inputs. Only the stored vectors determine the placement of decision boundaries that guide the classification process. Thus, the trained models are said to be \textit{sparse}. 

RVMs differ from SVMs in that they are based on Bayesian principles, which offer a number of advantages when simulating cognition and perceptual mechanisms. These include a probabilistic treatment of regression and classification and the potential to work with informative priors, and to compromise between model complexity and model accuracy.

Furthermore, the vectors which influence margin decisions in the RVM function as prototypes (hence ‘relevance’ vector) in that they usually do not appear close to class boundaries, but rather in regions of the feature space which are good representatives of the whole class population. The influence of prototypes on listeners' judgements of speech-sound category membership has been discussed in work on the perceptual magnet effect (Kuhl 1991) and its relevance to discrimination of phonetic detail within a phonemic category explored by Barrett (Barrett, 1997; Barrett-Jones & Hawkins, 2004).

Additionally, like support vectors, relevance vectors, while acting as prototypes, are actual observations coming from the training set, and so no averaging or smoothing of feature values needs to take place, since model sparsity is already achieved by vector selection. This feature allows phonetic detail to be retained and retrieved.

Finally, Tipping & Faul (2003) provide an algorithm for the RVM which performs incremental training and is thus particularly suited for simulating online learning. Hence it is relevant to perceptual learning by humans as well.

Taken together, these properties make the RVM a good candidate for a high-level, psycholinguistically appealing model.

Limitations of this technique are its static nature and expensive computation time during the training phase. This, however, is counterbalanced by much greater parsimony of the trained model compared with standard SVMs. Thayanantan (2006) presents a multi-class extension to the original binary classifier which is essential for applying the model to non-binary contrasts.

4.2 Specific properties of the current model

The version of the model tested here had several simplifications dictated by various constraints, most notably the limited nature of the input data and the early stage of development. Nonetheless, the key assumptions were retained of a Bayesian approach to the learning problem, and capability for both peripheral perceptual processing and higher level learning. These seem basic to achieving the desired balance between episodic and abstract aspects of classification, in achieving our main aim of mapping acoustic-phonetic detail onto morphological rather than segmental or phonemic identity.

The simulations reported here all use Gaussian kernel functions. For RVM training, no development set was employed to drive the choice of parameter values, for two reasons. First, a development set would reduce the size of the already small training and test sets. Second, preliminary tests showed that the only free parameter that is not adjusted automatically during the training process, the kernel width for the RVM, produced good classification performance when set to the mean standard deviation for all vectors in the training set. Since our goal was not to maximise classification accuracy, but rather to compare the relative contribution of various feature sets to the model accuracy and sparsity, these settings seemed to be optimal, at least pending new data about the discrimination performance of listeners.
5. SIMULATIONS

We assessed: A) the suitability of the RVM to the problem and the value of a sparse vs. a less sparse, non-Bayesian model; B) the convenience of the auditory primal sketch as opposed to a standard measure of energy; C) the benefits of using the cochlear model as a first filtering stage to the auditory primal sketch.

5.1 Simulation A: Relevance Vector Machine

5.1.1 Motivation

Baker et al. (2007a) showed that systematic phonetic differences between true and pseudo prefixes are perceptually salient: when a stem is preceded by a first syllable of the wrong morphemic type (for example when dis- from discover is spliced onto -colour from discolour), then the word is less intelligible in noise than when the stem is preceded by a cross-spliced syllable of the right morphemic type (e.g. when dis- from one token of discolour is spliced onto -colour from another token of discolour). The RVM can model distributions of partially overlapping data, and the Bayesian principles underlying the RVM offer good generalisation capabilities. Discrimination between the two prefix categories was therefore expected to be better than chance. Perfect classification was not expected, mainly because the acoustic distinction between the categories is somewhat gradient: aspects of the distributions can overlap in natural speech, with certain pragmatic contexts reducing the differences between some of the acoustic parameters that distinguish words with pseudo and true prefixes. Further, human ability to discriminate between these syllables is probably enhanced by hearing sufficient context to determine speech rhythm. Applying the model to isolated syllables was thus a stringent test of the discriminability of the syllables, and a better-than-chance result is all that was needed at this early stage of modelling.

Since the RVM displays properties that are appealing for the modelling of aspects of human speech recognition, we tested its performance against that of an SVM classifier trained on the same data. Performance metrics are discussed in section 5.1.3.

5.1.2 Research questions

Three research questions were asked in this stage of the modelling. 1) Does a sparse classifier achieve better than chance performance on the classification task? 2) Does an RVM classifier achieve higher classification accuracy than an SVM one on the same data? 3) Does RVM achieve greater sparsity than SVM?

5.1.3 Method

The simulation comprised training and testing RVM and SVM classifiers on the manual segment durations ([m]/[d], [t], [s], stop closure, VOT) of the stimuli as measured by Baker (2008, Baker et al. 2007a) and described in Section 3. Each of the 5 speaker-dependent classifiers was trained and tested with 5-fold cross-validation.

For each test set, accuracy was measured as area under the receiver operating characteristics curve (in short: area under the curve, AUC). The use of AUC as opposed to standard accuracy for measuring machine learning algorithms performance has been advocated by Bradley (1997), as it gives a more accurate picture of model performance, which is independent from the decision thresholds and cost functions adopted. For example, AUC is robust to changes in the prior probability of one category with respect to the other.
AUC for each speaker dependent classifier was averaged over the five train/test splits. This gave five AUC scores respectively for RVM and SVM. A mixed-effects ANOVA with RVM vs SVM as fixed factor and speaker as random factor was then run to highlight any significant differences between RVM and SVM accuracy.

A second measure of system performance was sparsity, characterised as discussed in section 4.1.2. Sparsity was measured as the number of training vectors retained by the model to build the mapping function of the classifier, and thus to make decisions about category membership of test samples. The smaller the number of vectors (decision vectors, DV) retained, the sparser the resulting classifier. As with the AUC, five measures were obtained for RVM and SVM, each of them averaged over the five train/test splits. Significance of difference in sparsity was tested with a mixed-effects ANOVA, with RVM vs SVM as fixed factor and speaker as random factor.

For SVM, optimal parameter values for $\gamma$ and $c$ were obtained with a grid search. For RVM, as mentioned, the only parameter that required an explicit setting (kernel width) was set to the mean standard deviation of the training data set.

5.1.4 Results

Detailed results are given in Figure 2. With accuracy rates between about 80% and 99%, both RVM and SVM models perform well above chance (50%). The slight but consistently better performance of the SVM just misses statistical significance ($F(1,4) = 7.31, p = 0.053$). In contrast, the RVM is significantly more sparse than the SVM ($F(1,4) = 34.42, p = 0.004$).
5.2 Simulation B: Auditory Primal Sketch

5.2.1 Motivation

Simulation B assesses the applicability of a model of rhythm based on general auditory processes to the *true vs. pseudo* morphological distinction in these syllables. At issue is whether the model captures small but systematic differences in the internal acoustic structure of similar acoustic segments, which together signal the difference in morphological structure.

Measures that automatically capture the absolute and relative durational properties of the acoustic segments (and thus may simulate aspects of perceptual integration) are good candidates for capturing the morphological distinction of interest. They should discriminate with above-chance accuracy comparable to levels achieved in Simulation A.

As discussed in section 4.1.1, the auditory primal sketch provides a combined measure of relative amplitude and duration of an event with respect to other neighbouring events.

In this simulation, it was compared to a vector which represented signal energy. Both kinds of vectors give a measure of amplitude modulation over time.

5.2.2 Research questions

The kinds of questions are essentially the same as for RVM vs. SVM, i.e.: 1) Do automatically extracted features achieve better than chance performance on the

---

### Classification accuracy and sparsity of RVM and SVM

<table>
<thead>
<tr>
<th>Speakers</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F(1,4)=7.31, p = 0.053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RVM</td>
<td>.946</td>
<td>.964</td>
<td>.879</td>
<td>.876</td>
<td>.819</td>
</tr>
<tr>
<td>SVM</td>
<td>.974</td>
<td>.974</td>
<td>.978</td>
<td>.939</td>
<td>.970</td>
</tr>
<tr>
<td><strong>DV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F(1,4)=34.42, p = 0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RVM</td>
<td>18.6</td>
<td>59.0</td>
<td>102.2</td>
<td>104.0</td>
<td>73.4</td>
</tr>
<tr>
<td>SVM</td>
<td>22.4</td>
<td>20.8</td>
<td>21.8</td>
<td>12.6</td>
<td>48.2</td>
</tr>
</tbody>
</table>

Figure 2: Simulation A: classification accuracy and sparsity of RVM and SVM. *Top:* area under the curve (AUC): accuracy. *Bottom:* number of decision vectors (DV): sparsity. Each of S1...S5 bar charts represents a model trained on a single speaker.

All values averaged over 5 train/test splits.
classification task? 2) Does an APS-based vector achieve better classification accuracy than an energy-based vector? 3) Does an APS-based vector achieve greater sparsity than an energy-based vector?

5.2.3 Method

Classification was carried out by an RVM. The APS vector was composed of the events automatically extracted by the auditory primal sketch with cochlear model. Each event was represented by a single number which indicates the prominence of the event with respect to the other events in the vector. The higher an event, the higher the number (see Fig. 1).

Because of the irregular number of events detected by the algorithm for each audio file, all vectors were aligned according to the most prominent event, and padded with zeros as necessary. Although this was our best option at the time, and a quite reasonable one given the nature of the data, this solution has greatly penalised the APS in this particular task. A new solution has since been developed. It consists in using a smooth version of the APS, which can then be sampled with a constant number of frames. This measure will be adopted in future evaluations.

The energy vector was composed of 10 frames of energy sampled at regular intervals and compressed by a factor of 0.3 in order to simulate psychoacoustic loudness perception curves (Stevens 1955).

![Classification accuracy and sparsity of APS and energy](image.png)

Figure 3: Simulation B: classification accuracy and sparsity of APS vs. energy. Top: area under the curve (AUC): accuracy. Bottom: number of decision vectors (DV): sparsity. Each of S1...S5 bar charts represents a model trained on a single speaker. All values averaged over 5 train/test splits.
5.2.4 Results
Detailed results are given in Figure 3. With accuracy rates between about 59% and 92%, both APS and energy vectors perform above chance (50%). The consistently better performance of the energy vector is statistically significant ($F(1,4) = 11.94, p = 0.025$). In contrast, differences in sparsity are not significant ($F(1,4) = 0.02, p = 0.90$). Whereas for AUC the direction of the difference is the same for all speakers, for DV it is not.

5.3 Simulation C: Cochlear Model

5.3.1 Motivation
Models of peripheral auditory processing present several advantages, such as a simulation of neural onset enhancement and inhibitory effects, and a perceptually-motivated representation of amplitude distribution across frequency channels. These features should help in the formation of complex auditory objects, and in the segmentation of the continuous waveform into perceptual units of various sizes. We wanted to see if the use of a simple cochlear model would have a positive impact in the extraction of auditory events by the auditory primal sketch. What was ultimately expected from the use of the cochlear model was a reduction in random variation, and a subsequent improvement in terms of accuracy and sparsity.

5.3.2 Research Questions
Even in this simulation, our research questions followed the pattern of the previous simulations. 1) Do cochlear model-filtered (CM) vectors achieve higher classification accuracy than non cochlear-model ones (NCM)? 2) Do CM-filtered vectors achieve greater sparsity than NCM?

5.3.3 Method
Two kinds of vectors based on the auditory primal sketch were compared in the classification task using an RVM classifier. The CM vector was based on the auditory primal sketch taken on the output of a cochlear model (Meddis 1986; Patterson et al. 1988). The NCM vector was based on the auditory primal sketch taken on a downsampled version of the absolute (rectified) of the waveform. In both cases, each event was represented by a number as described in sections 5.2.3, 4.1.1 and Figure 1.

5.3.4 Results
Detailed results are given in Figure 4. The consistently better performance of the APS with cochlear model (CM) is statistically significant ($F(1,4) = 12.27, p = 0.025$). Furthermore, the CM vector is significantly more sparse than the NCM one ($F(1,4) = 54.49, p = 0.0018$). The direction of the difference is the same for all speakers, both for AUC and DV.
6. DISCUSSION

We have outlined a model architecture which accounts for the role played by phonetic information in the specific case of bi-syllabic words whose first syllables are transcribed with the same phonemic sequence, but which differ in acoustic-phonetic detail that varies systematically with morphological status. The main model assumptions include a Bayesian view of the learning and recognition process, the centrality of low-level perceptual representations, and hybrid episodic-abstract high level representations: that is, a representation which neglects phonemic identity, or at least does not give it greater importance than morphological identity.

The simulations presented here are encouraging as to the model components selected, both in terms of accuracy in discrimination tasks and explanatory elegance. These components are now being integrated into an architecture that will allow model performance to be compared with perceptual data.

To the best of our knowledge, this is the first time that prominence scores based on Todd's (1994) model have been used in a learning setting for speech. RVMs have been used in a speech recognition task in at least one other case (Hamaker et al. 2002).

To implement a fuller speech recognition system based on these principles, several extensions are required, most notably a way to apply the high level model to longer

---

**Figure 4: Simulation C: classification accuracy and sparsity of APS with cochlear model (CM) vs. APS without cochlear model (NCM). Top: area under the curve (AUC): accuracy. Bottom: number of decision vectors (DV): sparsity. Each of S1...S5 bar charts represents a model trained on a single speaker. All values averaged over 5 train/test splits.**

<table>
<thead>
<tr>
<th>AUC F(1,4)=12.27 p = 0.025</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
</tr>
<tr>
<td>CM</td>
<td>.756</td>
</tr>
<tr>
<td>NCM</td>
<td>.646</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DV F(1,4)=54.49 p = 0.0018</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
</tr>
<tr>
<td>CM</td>
<td>43.8</td>
</tr>
<tr>
<td>NCM</td>
<td>69.6</td>
</tr>
</tbody>
</table>

n = 160
sequences of variable duration and number of sub-units (acoustic segments). More low-level perceptually motivated features are also needed: prominence information based on duration and intensity provided by the rhythmogram should be complemented with other cues based on spectral information (timbre and formant structure) and possibly relative pitch.

Although our work has focused on using acoustic-phonetic detail to make a specific morphological distinction, the general principles of the model are expected to apply to other comparable distinctions.

ACKNOWLEDGEMENTS

Funded by EU MRTN-035561 Sound to Sense. We thank Rachel Baker for sharing her data and analyses, and members of Sound to Sense for guidance and discussion.

REFERENCES


Bradley, A. (1997), The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30(7), 1145-1159.


