Automatic recognition of repetitions and prolongations in stuttered speech

Article · January 1995

2 authors, including:

Peter Howell
University College London
268 PUBLICATIONS **6,923** CITATIONS

AUTOMATIC RECOGNITION OF REPETITIONS AND PROLONGATIONS IN STUTTERED SPEECH

Peter Howell and Stevie Sackin

This contribution reports the first attempt at training artificial neural nets to locate stutterings. The particular stuttering events to be located are repetitions (reps) and prolongations (pros). There are undoubdtedly problems in characterizing stuttering on the basis of these and other canonical dysfluency categories and for this reason it is anticipated that a move will be made in the future to other (acoustical/non-canonical) criteria. However, these events have been selected for investigation at this stage of our investigations for several reasons. The principal one is that reps and pros are ubiquitous in the speech of stutterers (Wingate, 1988) and as a result of this are widely used by clinicians. Artificial neural networks can be trained to learn automatically what information in suitably-coded input presented to it differentiates events of one type from others.

Recognition by an artificial neural network is achieved here by marking which words are repeated, which prolonged and all others (including fluent speech and "other"dysfluency categories). Separate artificial neural networks are trained to differentiate either reps or pros from all remaining speech. The training is achieved by linking the acoustic input representations either onto rep or pro outputs (as appropriate) through a layer of hidden units (Rumelhart and McClelland, 1986). During this training phase, the weightings of the network are automatically adjusted to bring the mapping between input and output into correspondence. Thus, an artificial neural network that has learned to recognize a rep would give a high output during a stretch of speech that contains a rep and a low output elsewhere. Explicit definition of the way that acoustic information is mapped onto dysfluent events is not needed. Once trained, the artificial neural network will classify the reps and pros in speech recordings not encountered before. The main issues in getting the artificial neural network to learn the mapping between acoustic structure and the canonical dysfluency categories reps and pros are, 1. How to select training data that judges can agree on, 2. The appropriate form of the acoustic input, 3. The topology to use for the artificial neural network, 4. Evaluation of whether the goal of recognizing the events has been achieved.

1. Training data

This issue is inherently linked with obtaining training instances which judges agree about stuttering category: How can you select data to train an artificial neural network if judges can't agree on whether a particular episode is a rep or pro? The technique we have developed is described in detail in Howell and Sackin (in press) and in outline in Howell, Sackin and Cowan (this volume). Each word was judged as fluent, rep, pro, and all other dysfluency categories given by Wingate (1988). For each of these categories (including the fluent category), the judge gave a rating for each word reflecting where the word lies in its category relative to the extremes of the fluent and dysfluent categories (1= far from the dysfluent extreme, 3 = adjacent to fluent extreme and 2 = intermediate).

2. What is the appropriate form of the acoustic input?

The choice of the appropriate acoustic input depends on what sort of structure reps and pros

have that distinguishes them from other speech with repetitive or prolonged tendencies. The choice of acoustic pre-processing should then seek to enhance these properties.

Two alternative acoustic representations which highlight this structure have been investigated in preliminary studies (see Howell, this volume for an outline of other properties being investigated for locating repetitions): (a) A combination of the autocorrelation function and spectral information and (b) Envelope parameters.

(a) The autocorrelation function and spectral information

The autocorrelation function (acf) is a correlation of the speech signal with a time-shifted version of itself. A particular point in time can be correlated with points preceding and following, (only correlations later in time are looked at here). Typically, in speech analysis, time shifts of a few ms are employed, which allows the periodicity in, for example, pitch pulses in voiced speech to be measured. If there are sections which are repeated or sustained for long periods of time (reps and pros), the acf coefficients will be high as long as the time interval is within that section. Conversely, time shifts at which the acf coefficients are high indicate that a component is repeating or sustained at that periodicity. The events to be detected here are either sustained or repeated at rates longer than the duration of speech segments. To locate reps and pros, the speech is split into adjacent frames of 10 ms duration. For each frame, an acf is obtained for 20 time shifts ranging from 100 ms to 1050 ms, in 50 ms steps. These time shifts were chosen to encompass regions of reps and pros. The analysis is performed at successive points in time (as in spectrographic analysis), so repeated or prolonged sections extending over greater periods of time are represented over successive frames.

A plot of the acf for different time shifts at a point within a rep or pro allows these events to be differentiated from each other. In the pro, the acf is high for a broad range of time shifts. In the rep, the acf for different time shifts is multilobed at time shifts corresponding specifically to multiples of the repetition rate of the attempts. Departures from regular timing in reps are evidenced in somewhat broader tuning of the peaks. Mixtures of reps and pros (e.g., amplitude modulations on sustained sounds) occur and contain features of both dysfluent event types. Thus, the coefficents are high throughout the event but still contain discernible peaks.

For all 10 ms frames within a rep or pro, there are, then, 20 coefficients at any point in time. The two patterns that characterize reps and pros extend along adjacent frames.

Spectral inputs are, at present, inputs from a 19 channel vocoder.

(b) Envelope parameters

A considerably simpler representation of the acoustic structure which retains information about reps and pros is the envelope. The envelope of a speech waveform was obtained by full wave rectifying the signal and low-pass filtering 10 Hz. The envelope is a simpler representation than acf plus spectral coefficients since there is only a single coefficient (amplitude envelope value) at any point in time.

3. What topology to use for the artificial neural network?

Separate artificial neural network models are trained for recognizing reps and pros, at present.

(Once full investigations have been carried out, an integrated artificial neural network will be constructed for recognizing reps and pros separately within the same model.) Currently, then, the models used for learning the structure of stuttered speech consist of a set of vector inputs which map each frame onto a binary output (e.g., rep or not) through a layer of single hidden units. The artificial neural network is trained by presenting acoustic vectors of the category to be identified at input, and any identification error at output is back-propagated with weights being adjusted at the hidden unit layer so that the vector at the input maps onto the required output. The steps that remain to be specified are how the artificial neural network is configured (**model specification**) and the options available in the **training software**.

Model specification. In our implementation, the topology of the artificial neural network is specified in a text file. This contains several items of information. Firstly, the input to the artificial neural networks is in the form of vectors which are obtained from the acoustic analysis software which can span a single or odd multiple frames. These can be single types of input (e.g., envelope parameters) or mixed inputs (e.g., acf and spectral coefficients). Secondly, besides providing information about vector input, the way the network is interconnected can be specified.

The vector input to the artificial neural network is the 20 acfs plus 19 vocoder values (39 in total) in a frame or these vectors for an odd multiple of the number of frames. (The reason that an odd multiple is needed is so that the window has a central value which makes it symmetric in time). The vector input from the envelope analysis would be the output of a single or odd number of frames. The higher the multiple, the more frames are included in training. Thus if information over three frames (30 ms) is included, 117 inputs are specified for acf plus vocoder vectors.

The net is at present fully interconnected. The spectral inputs from the vocoder may allow the network to inhibit a prolongation response to fluent vowels. If a vowel inhibitor is to be modeled explicitly some of the low frequency spectral inputs would be exclusively connected to certain hidden units to provide low outputs when pro output is high and vice versa. This would provide competing patterns: High vowel output on the vowel inhibition part would be associated with low pro output and vice versa. Alternatively, the artificial neural network should learn to do this without explicitly specifying the inhibitory connections which can be confirmed by examining the weights on the hidden units. Prolngations on vowels would be suppressed too but there are few of these in adults speech according to Brown (1945).

4. Evaluation of whether the goal of recognizing the events has been achieved.

Once a distinction has been satisfactorily learnt, this can be applied to novel test data. To perform this, examples of non-stuttered speech and each type of dysfluent event are needed.

The running of the artificial neural network on unfamilar data uses the same software and input as during training, but two changes should be noted. First, a flag implies that the weights presented should be used (and the weights should not be adapted further). Second, vectors for the data to be tested have to be provided.

5. Results and applications

Preliminary results have been obtained from an artificial neural network trained with 20 acf and 19 vocoder coefficients/10 ms frame and with twenty frames of one envelope coefficient/10 ms frame as input with separate nets for reps and pros (four in all). (Differences in severity were not included as a factor in training.) All networks were trained with 2 min. of speech from the same adult speaker and tested against this speech and that of five further speakers. The artificial neural networks were scored against data in which the reps and pros were differentiated by the judges as severe, moderate or mild (using the ratings given to these types of dysfluency). Hit/miss rate was 0.82, 0.74 and 0.57 for pros and 0.77, 0.71, 0.55 for reps using acf-plus-spectral coefficient input. Hit/miss rate was 0.79, 0.67 and 0.55 for pros and 0.71, 0.65 and 0.53 for reps using envelope parameters alone. Note the gradation of performance which duplicates 8human performance (Boehmler, 1959). Also, although the performance is somewhat poorer for the envelope coefficents, the envelope input is simpler than the acf-plus-spectral coefficient input. Note that there is still scope for improvement, particularly in the case of repetitions.

Acknowledgement: This research was supported by a grant from the Wellcome Trust.

References

Boehmler, R. M. (1959). Listener responses to non-fluencies. Journal of Speech and Hearing Research, 1, 132-141.

Brown, S. F. (1945). The loci of stuttering in the speech sequence. Journal of Speech Disorders 10, 181-192.

Howell, P. and Sackin, S. (in press). Acoustic assessment of stuttered dysfluent speech. Chapter to appear in: "Current Topics in Acoustical Research".

Rumelhart, D. E. and McClelland, J. L. (1986). Parallel-distributed processing: Explorations in the microstructure of cognition: Vol 1. Foundations. MIT Press: MA.

Wingate, M. E. (1988). The structure of stuttering: A psycholinguistic study. Springer-Verlag, 1988.