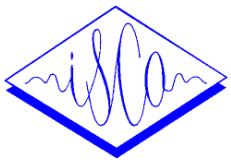

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

AN ARTIFICIAL NEURAL NETWORK FOR THE BURST POINT DETECTION

Shigeyoshi Kitazawa and Masahiro Serizawa

Shizuoka University
Faculty of Engineering, Department of Computer Science
Hamamatsu 432, Japan

Abstract

We trained a three-layered neural network to discriminate between the left and right contexts around the burst point. The shifting window is the portion of the input speech which will serve as input to the network at a time. The detection window is the portion of the input speech where the searched burst locates. The outputs from our network distinguish three states, i.e. before or after the burst and outside the burst of the shifting window.

The acoustic properties used was speech power time series of a 5-band mel-scaled LPC spectrum. The network consisted of 80 input units 30 hidden units and 3 output units. French voiced stop consonants /b, d, g/ served as input. Elimination of phantom burst point was attained. The decisions error was around 15-ms. Usually right decisions consistently proceeded the real burst point and then followed by left decisions. High frequency component was effective among 5 frequency bands.

I. INTRODUCTION

The object of this research was to extend the burst point alignment to burst point detection task, focusing on ways to train a connectionist burst point detector. We were motivated in this direction by two factors:

- Issues of speaker independence.
- Suppression of phantom burst point.

We wanted our model to have the flexibility necessary for robust recognition across a range of speakers. There were difficulties in the connectionist approach in applying to rather complicated problems. There were a few reports on recognition of speaker independent systems. Speaker independent network was very difficult to accomplish because:

- Even with the speaker dependent mode, the connectionist network takes redundant connection network in order to converge to solutions. The order of network is much larger than linear discriminant functions which adjust many weights during learning by backpropagation errors.
- In order to cover virtually all possible speakers, the supposed number of speakers was around several hundred or several thousands or more, which implies that the connectionist network requires huge database.

We challenge to this speaker independence by avoiding above difficulties. We simply distinguish two states. We started from a small number of speakers but included all possible combination of consonant-vowel syllables. Comparison with linear discriminant function was in our mind. With its robust characteristics guided performance prediction and feature selection.

The burst point is the point of time where the plosion begins. Plosives are classified into voiced/voiceless or bilabial/dental/velars (there is also a few possibility of burst in initial vowels such as glottalized stops). Here we focused on detecting the weak and unclear burst of voiced stops. We considered this problem with interest in application to speech recognition. Elimination of false alarm is important figure of performance as well as correct detection. To help this we used three output units including the suppressing output.

Why the problem of the burst point detection is important in speech recognition? First, the invariant feature resides near the burst point. We have shown spectral change around the burst point determined the consonant place of articulation independent of vowel context, where we hand labeled the burst point. [1] Replacement with some automatic algorithm has been a theme for the further development. Second, phonological perception comes up with a set of basic features such as distinctive features before reaching at phonetic recognition. With this respect, the burst point relates to one of the major features consonantal. Third, the feature of stop consonant resides in the burst envelope as well as spectrum because the envelope determines the burst point. Fourth, recent analysis show that the burst like change in waveform is remarkably enhanced in the auditory system, therefore the burst point should be a significant feature for human perception.

In the following sections, we discuss what is the feature of the burst point, why we adopted the artificial neural network (ANN), and how we could suppress number of

false alarm. Finally, we will show about the further possibility of the network.

II. BURST DETECTION

2.1 Burst point features

There are several features used for burst point detection as follows. They are for segmentation, spectrogram reading, and speech recognition.

(0) raw waveform,
(1) waveform envelope,
(2) high frequency energy,
(3) thin line on a sonagraph,
(4) discontinuous spectral change, and
(5) direct phoneme detection.

One of approach using the first feature was pattern recognition of simplified the waveform representation[2]. The waveform envelope implies local and global changes. In voiced stops, a small wedge-shaped spike appears along with the prevocalic vibration. A sharp onset of the following vowel is an indirect feature of stop burst. At the burst point, high frequency component rise up suddenly, though the power is not so high. Spectrogram readers describe thin line after the closure as an indication of stop consonant. It is an intuitive prospect that short term spectrum discontinuously change at the burst point from steady state. Another possibility is direct detection of phoneme by shift invariant features, for example by neural networks or HMMs[3, 4]. The last method extracts burst feature implicitly integrating the transient feature. This usually requires large network and large training data and large computation. Therefore this is still difficult to extend to speaker independent mode.

Every feature closely related with the method appropriate, i.e. signal processing, pattern recognition, knowledge representation. Here we focused on the waveform feature, because it is independent from spectrum and important perceptual cue. We recognized limitations of spectral features from our experience in statistical analysis of time varying spectrum[1]. At the same time, we also recognized the potentials of envelop of several millisecond or tens millisecond as a feature through hand labeling of burst points.

The efficient representation of envelope is necessary, but we have less experience of this new features compared to spectral features. Thanks to all-powerfulness of the neural network, we skipped feature extraction and directly detected the burst point.

2.2 Architecture of the neural network

The architecture of the network employed was as follows. It was a multi-layered perceptron network with one hidden layer (an ANN). The number of the input unit was 80 and 30 hidden units and 3 output units.

The input units served as a window to observe input speech. The input data was a waveform envelope. When the observation window, i.e. 80 input units, scanned input speech from left to right, one of outputs fired. The output units corresponded to the following three states; the burst point was in the right half of the window or in the left half of the window and the burst point was outside of the observation window.

III. EXPERIMENTS

3.1 Speech analysis for inputs

Phonemes studying were French voiced stop consonants, i.e. /b/, /d/, and /g/. The data base consisted of monosyllables followed by 11 vowels spoken by 40 native male speakers as shown in Table 1. We choose burst point of initial position of speech, by careful observation of waveforms to find the earliest evidence of burst if there any multiple burst. In French, burst is very weak in energy, so this process was important to prepare teaching outputs. Digitization frequency was 16 kHz. Speech analysis procedure was as follows starting from the analysis point:

1. 256 point Hamming windowing,
2. 14 linear prediction coefficients,
3. 64 point LPC log spectrum,
4. log scaled 5 band power, and
5. 1 ms shift of the window.

The procedure repeated 80 times for each analysis point and then normalized for the input to the ANN. The analysis point moved from left to right so the burst point came in from right edge and went out from left edge of the observation window. The frequency axis divided in 5 power band from 0- to 1-kHz, 1- to 2-kHz, 2- to 3-kHz, 3- to 5kHz, and 5- to 8-kHz. Characteristics of each band are different.

3.2 Mechanism of burst point detection

The ANN used the back propagation algorithm DCP2 implemented in ATR. The network was a detector of the burst point. The 80 input units work as observation window for the input speech. It scanned input speech from left to right showing the status by three output units. The outputs include some possibility of errors. At the most possible burst point, steady right decisions followed by steady left decisions.

Table 1. French voiced stop syllables.

Table 4 Syllables for French Consonant Study											
Consonant	Vowel										
	[a]	[o]	[œ]	[ɛ]	[ɛ]	[u]	[y]	[i]	[ɑ]	[ɛ]	[ɔ]
[b]	ba	bo	beu	be	bai	bou	bu	bi	ban	bin	bon
[d]	da	do	deu	de	dai	dou	du	di	dan	din	don
[g]	ga	go	geu	ge	gai	gou	gu	gi	gan	gin	gon

3. 3 Training of the ANN

There are number of training patterns but we reduced training data in the following way. Although we intended the ANN to be speaker independent, we trained the network for four speakers. Each speaker pronounced 33 syllables (3 voiced stops and 11 vowels). The syllables amount to 132 from four speakers that separates two disjoint sets one for training and the other for test as shown in the Table 2.

The training set contains 66 syllables, however, training pattern amount to large numbers because shifted patterns near the burst are possible training patterns. For example, in 80 input units case, there are 80 possible input patterns that contain burst point inside. Therefore there are six thousand of possible input patterns. It is too much to consider all of these possible patterns. In order to reduce the number of training patterns we selected 30 input patterns out of 80 possible patterns, and added 4 outside the burst patterns. The relative position of training patterns are displayed in Figure 1. We prepared more patterns near the burst point to attain the finer precision, while fewer patterns offside the burst point from the observation window.

Each network for one of 5 power band separately adjusted weights by the back propagation algorithm with dynamic moment adjusting. After two or three thousand iterations, the networks converged to the recognition rate shown in Table 3..

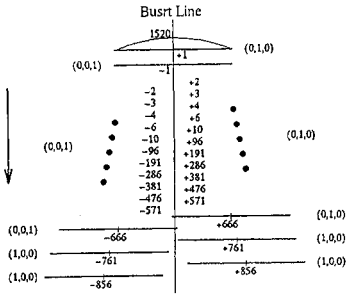

Figure 1. Training patterns prepared from each syllable.

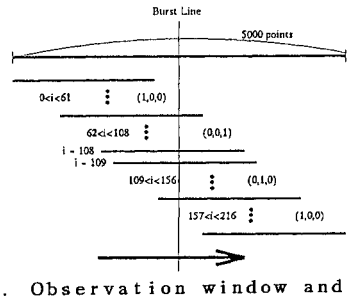

Figure 2. Observation window and expected output status.

3. 4 Recognition of the status

The absolute error rates of training data do not make sense unless the networks test the untrained data, but relatively lower performance in 1 and 4-state, i.e. the burst point is not within the detection window, was due to fewer training. Recognition experiment used the test data of the same set of speakers as well as the training data. The experiment was as follows:

(1) Recognition window was 2500 sample points before and after the burst. The analysis point began at the head of the window.
(2) From the analysis point, 80 input units entered to the ANN.
(3) The maximum output unit among three was the decision of the ANN.
(3) Shift the analysis point 1-ms and continue the procedure from (1). While staying within the recognition window, the observation window shifted 217 times.

The state and expected output change as in Figure 2 along the window shift. Difference between spectrum band was not clear from the training score, but the score on the whole of recognition window including untrained patterns showed significant difference of performance. Table 4. shows correct response rate for each interval of state on both learning syllables and testing syllables. There was not significant difference in performance between learning and testing data. The most useful spectrum band was 5-kHz over which included sudden build up of high frequency component of the burst. The second useful band was lowest frequency up to 1-kHz including build up of first formant component of the following vowel.

Table 2. Composition of data set for leaning and testing.

O: Training data X: Test data

Context	sp1000	sp1001	sp1002	sp1003	Total
a	O	X	O	X	4
o	X	O	X	O	4
eu	O	X	O	X	4
e	X	O	X	O	4
ai	O	X	O	X	4
ou	X	O	X	O	4
u	O	X	O	X	4
i	X	O	X	O	4
an	O	X	O	X	4
in	X	O	X	O	4
on	O	X	O	X	4
Total O	6	5	6	5	22
Total X	5	6	5	6	22

Table 3. Rate of correct classification of learning patterns in percent.

	frequency band in kHz				
decision	0-1	1-2	2-3	3-5	5-8
outside window	77	63	91	71	80
righthalf window	95	95	98	95	96
lefthalf window	97	95	99	95	98

Table 4. Correct classification rate for learning syllables and test syllables.

	/b/	/d/	/g/	average
trained	78	79	69	75
untrained	69	72	59	67

3.5 Comparison with discriminant analysis

One of classical approach is this discriminant analysis based on multi-dimensional normal distribution. The most robust method is linear discriminant functions which are available as "STEPDISC" and "DISCRIM" from the SAS Institute Inc..

The inputs were fifth band spectrum of 80 points of every milliseconds. The variables selected were 1, 5, 15, 23, 72, 80th among 80 input elements, since the others were highly correlated. The discriminant scores are in Table 5. which says about 50 to 60% of correct identification is possible for detection. The score achievable by ANN was comparable to that of linear discriminant function except small improvement. However, the identification score does not directly related with detection accuracy.

3.6 Burst point detection by the ANN

Frame identification made in the previous section was redundant because the output was possible at any analysis point. Therefore the burst point detection based on this sequence of outputs along time series of decision. The results shown in Figure 3 say high precision detection within a few millisecond error, quite low omission rate. The detection was unique in 95% of syllables, duplicated in 3%, and lost in 2%. The 1st and 4th state introduced worked effective to suppress redundant detections.

There were, however, several large errors of detected burst point which was due to difficult cases typically as shown in Figure 4 where envelope developed gradually from prevoicing through onset transition to the following vowel. The trained network seemed to be biased toward the trained speakers, since errors for untrained speaker superposed on the Figure 3 show rather distributed.

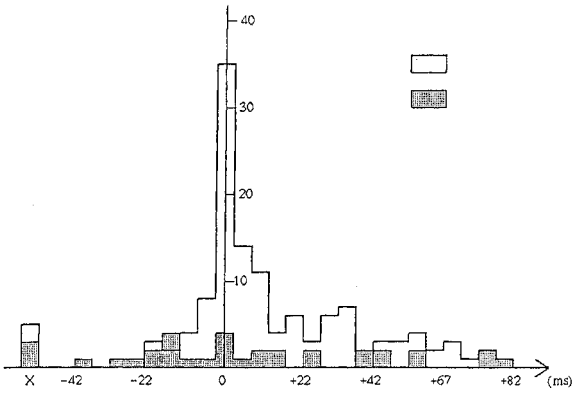

Figure 3. Histogram of errors of burst point detection. For the trained four speakers pooled in blank, and for an untrained speaker in shaded bars. The symbol 'x' show the cases where no burst point was detected.

IV. CONCLUSIONS

We confirmed previous experiments[5]. The informative frequency band was 5- to 8-kHz and 0- to 1-kHz, but 2- to 5-kHz was not so good as the other two. The detection rate was about 70% within 3-ms error. There was very few omissions. Training with additional state eliminated false alarms. ANN performed better than discriminant analysis.

Results were insufficient, which does not imply defects of ANN, but insufficient variation of training patterns, and number of hidden units and layers. Envelops of several tens milliseconds were the extracted feature which limited performance in case of gradual vowel onset for some /g/s. Short term envelope such as wedge-shaped spike is also a good feature. Some further improvement in signal processing and feature presentation is necessary.

Acknowledgments

We appreciate Mr. M. Nakamura, ATR interpreting telephony Co. Ltd., for supporting us with learning software.

References

[1] S. Kitazawa, J. P. Tubach, "Discriminant Analysis and Perceptual Test of French Stops and Nasals," 9th ICPR, Nov. 1988.

[2] S. Seidl, F. Poirier, "An Approach for Automatic Determination of Break Points in the Speech Waveform," Eurospeech'89, 2, 96-99, 1990

[3] J. B. Hampshire II, A. H. Waibel, "The Meta-Pi Network: Connectionist Rapid Adaptation for High-Performance Multi-Speaker Phoneme Recognition," ICASSP'90, S1, 165-168, 1990.

[4] S. Nakamura, K. Shikano, "Speaker Adaptation Applied to HMM and Neural Networks," ICASSP'89, S3. 3, 89-92, 1989.

[5] S. Kitazawa, M. Fourati, S. Ichikawa, "Burst point location in stop consonants using backpropagation neural networks," J. ASA Sup. 1, 84, S59, Nov. 1988.

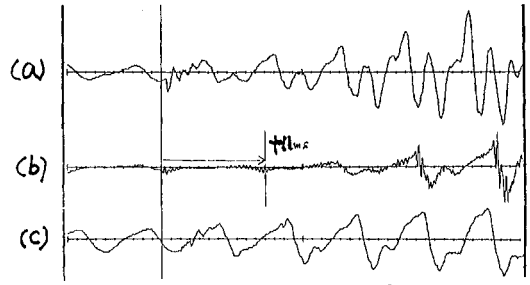

Figure 4. Example of waveform near the burst. (a) exact detection, (b) +11-ms error, and (c) unable to detect.

Table 5. Correct classification rate by the linear discriminant analysis.

	classified into (in percent)		
from	outside	right-half	left-half
outside	46	26	27
right-half	18	58	24
left-half	22	27	51