Neural Networks, IJCNN'05, IEEE Int. Joint Conf. on, Vol. 4, pp.2076-2080,
http://dx.doi.org/10.1109/IJCNN.2005.1556220

Exploration of Rank Order Coding with Spiking Neural Networks for Speech Recognition

Stéphane Loiselle and Jean Rouat
Département de génie électrique
et génie informatique
Université de Sherbrooke
Sherbrooke, QC, CANADA J1K 2R1
Stephane.Loiselle@usherbrooke.ca
http://www.gel.usherbrooke.ca/rouat/

Daniel Pressnitzer
École normale supérieure
Département d'Etudes Cognitives
45 rue d'Ulm
75230 Paris Cedex 05
France
Email: Daniel.Pressnitzer@ens.fr

Simon Thorpe
Centre de Recherche Cerveau et Cognition
Faculté de Médecine de Rangueil-Bât A3
133 route de Narbonne
31062 Toulouse Cedex 05
France
Email: simon.thorpe@cerco.ups-tlse.fr

Abstract— **Speech recognition is very difficult in the context of noisy and corrupted speech. Most conventional techniques need huge databases to estimate speech (or noise) density probabilities to perform recognition. We discuss the potential of perceptive speech analysis and processing in combination with biologically plausible neural network processors. We illustrate the potential of such non-linear processing of speech by means of a preliminary test with recognition of French spoken digits from a small speech database.**

I. Introduction

Statistical approaches like Bayesian networks and Hidden Markov Models perform reasonably well only when the probability distributions have been suitably estimated during the training phase. State of the art speech recognizers first extract parameter features characteristics of speech signal (analysis module) and then classify (recognition module), in the feature space, the input vectors.

The classification is done by comparing the input sequence vectors with pre-stored models of each word (or sub–word unit) of the vocabulary [1]. The speech analysis module usually extracts every [10–35] ms a vector of Mel Frequency Cepstrum Coefficients (MFCC) [2]. A Markov model is associated to each sub-word unit in the dictionary. First order Markov chains with Gaussian mixtures are commonly used. Gaussian mixture parameters (mean, covariance matrices, mixture coefficients) and transition probabilities of the Markov chains have to be estimated during training [1], which usually requires supervised training using huge hand-labelled databases [2] and thus these speech recognizers are designed for specific applications. In fact, for robust speech recognition (noisy speech and interfering noise), these systems have to be also trained on noisy databases where different combinations and configurations of noise or interference should be present.

On the other hand, perceptive and bio-inspired approaches require less training, can be unsupervised and offer strong potential even if their use in speech recognition applications are less mature. In the present work we are interested in a monophonic bio-inspired approach for speech recognition with limited training. From preliminary experiments, it is observed that this kind of approaches could be a good complement to statistical speech recognizers as they might reach acceptable results more quickly on very limited training sets.

II. Exploration in Speech Recognition

In pattern recognition research, it is well known that signal analysis and recognition are modules that are tightly related. For example, very good matching between parameter vector distributions (such as MFCC) and recognition models (such as HMM) yields better performance than systems using auditory cues but with mismatched pattern recognizers. Further discussion is given by M. Hunt in [3], [4].

Research in neuroscience and auditory perception has advanced, yielding greater understanding of the auditory system along with more sophisticated tools for the recognition of time-organized features. See for example the work by Zotkin et al. [5]

We illustrate here an application in speech recognition where perceptive signal analysis combined with non-linear signal processing and spiking neural networks offers a strong potential. Bio-inspired neuronal networks are well adapted to signal processing where time is important. They can be fully unsupervised. Adaptive and unsupervised recognition of sequences is a crucial property of living neurons. At the moment, this work does not reflect the full potential of spiking neurons and is more or less exploratory.

[1] Word or sub–word units are syllables, phones, phonemes, etc. depending on the approach in use.

[2] MFCC computation:

1) Signal preaccentuation;
2) Hanning windowing;
3) FFT and module extraction;
4) bank of Q-constant filters (usually 24 filters) and extraction of the filterbank output energies;
5) Logarithm of these energies.
6) Discrete Cosine Transform of the log energies.

A. Speech Recognition with Ensemble Interval Histograms

Oded Ghitza proposed in 1994 and 1995 the use of an auditory peripheral model for speech recognition [6], [7] that simulates a great number of neurons with different internal threshold values. O. Ghitza introduced the notion of the Ensemble Interval Histograms representation (EIH). That representation carries information about the spiking time interval distributions from a population of primary auditory fibers. Experiments were made on the TIMIT database by using a mixture of Gaussian Hidden Markov Models. He observed that the EIH representation is more robust on distorted speech when compared to MFCC. On clean speech there were no gains in using his model.

It is important to note that EIH carries information on averaged spiking intervals, thus specific sequences of spikes cannot be identified inside a population of neurons. Furthermore, the representation has to be smoothed to be compatible with the use of a conventional fixed frame pattern recognizer (HMM with multi-Gaussian). Therefore, fine grained information is lost. There is, however, increasing evidence that this information could be used by the auditory system [8], [9].

We suggest to use a similar front-end as proposed by Ghitza, but to preserve the time structure organization of spiking sequences across neurons, without computing the histograms. As this approach prevents the conventional use of HMM, we examine potential techniques to recognize specific spiking sequences. Different coding schemes can be used to perform the recognition. Here we investigate the Rank Order Coding scheme (ROC). The ROC scheme has been proposed for visual categorization by Thorpe *et al.* [10], [11]

B. Rank Order Coding

Rank Order Coding has been proposed by Simon Thorpe and his team from CERCO, Toulouse to explain the impressive performance of our visual system to recognize objects [12], [13]. The information is distributed through a large population of neurons and is represented by spikes relative timing in a single wave of action potentials. The quantity of information that can be transmitted by this type of code increases with the number of neurons in the population. For a relatively large number of neurons, the code transmission power can satisfy the needs of any visual task [12]. There are advantages in using the relative order and not the exact spike latency: the strategy is easier to implement, the system is less subject to changes in intensity of the stimulus and the information is available as soon as the first spike is generated.

Perrinet [13] and Thorpe [11] also discuss the importance of sparse coding and rank order coding for classification of sequences which is particularly interesting for speech. R. VanRullen *et al.* also discuss findings for spike-time coding and their potential generalization to sensory modalities, including the auditory system. [14]

III. System Overview

We now explore the feasibility of a speech recognizer based on spiking neurons and rank-order coding. In this work, we use the temporal order of spikes within peripheral auditory channels as the basis for the speech recognition. Spikes are obtained with simple neuron models as fixed input thresholds, without integration. Different neurons can have a different threshold, but the leak and the integration from the conventional Leaky Integrate and Fire neuron model are not used in this first implementation.

A. Speech Analysis Module

The peripheral auditory system is crudely modeled by a gammatone filter-bank [15] followed by rectification and compression with a square-root law. The output of such a simulation is intended to represent the average firing probability in primary auditory nerve fibers, without taking into account adaptation and loss of phase-locking. For each channel, we then use three neurons with different thresholds (denoted 1, 2 and 3). The firing probability is converted into actual spikes when it exceeds a neuron's threshold. After producing a spike, a neuron becomes inactive for the remaining time of the stimulus.

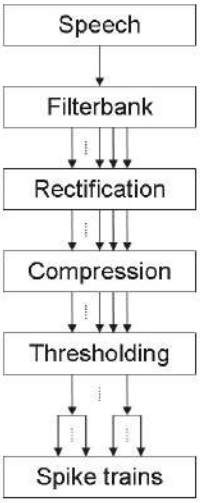

Fig. 1. Speech Analysis Module. The signal is filtered by a cochlear gammatone filter-bank, then each output is rectified and compressed. Finally, spikes are generated by multiple thresholding in each channel.

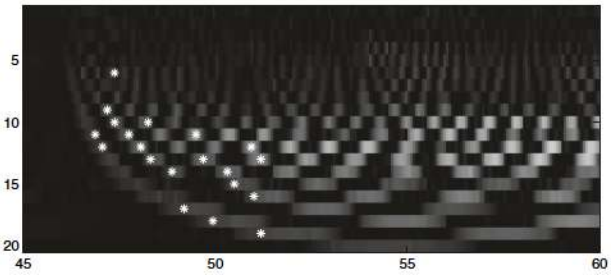

Fig. 2. Spike train generation illustrated on a French digit 'un'. Spikes (stars) are generated when the signal (French digit 1) amplitude exceeds the thresholds. The x-axis represents the samples (sampling frequency of 16 KHz) and the y-axis shows the filter-bank channels. Center frequencies of channels 1 and 20 are respectively equal to 8000 and 100 Hz.

B. Learning and Recognition Modules

During learning a template is generated for each reference word. The template consists of the sequence of N cochlear channels indices that are most likely to be activated first. As an example, with the French digit "un" pronunciation (figure 2), the sequence of twenty channels (table I, columns **Position**) is : 11, 12, 9, 10, 6, $11_{(2)}$, $12_{(2)}$, $10_{(2)}$, 13, 14, 17, $11_{(3)}$, $13_{(2)}$, 18, $14_{(2)}$, 15, $12_{(3)}$, 16, $13_{(3)}$ and 19, where the (i) indices stand for threshold levels 2 or 3 in the corresponding channel.

TABLE I

POSITION IN THE SEQUENCE OF THE FIRST 20 CHANNELS TO PRODUCE A SPIKE AND THE GENERATED WEIGHTS

	Position			**Weight** $(k_{(i)})$		
	Levels			Levels		
Channel	1	2	3	1	2	3
1						
2						
3						
4						
5						
6	5			16		
7						
8						
9	3			18		
10	4	8		17	13	
11	1	6	12	20	15	9
12	2	7	17	19	14	4
13	9	13	19	12	8	2
14	10	15		11	6	
15	16			5		
16	18			3		
17	11			10		
18	14			7		
19	20			1		
20						

Afterwards, weights k (column **Weight**, table I) are given to each of these channels according to Equ. 1.

$$k_i = (N - i) + 1, \tag{1}$$

where i takes values from 1 to N. As we can see, the weight change depends on the channel position in the sequence. The first channel to generate a spike has the highest weight and we keep going downwards until we reach the N^{th} channel. Since there are more than one neuron per channel, more than one weight can be given to a single channel (see table I).

To perform isolated word recognition, signals of words to be recognized are filtered, rectified, compressed and thresholded (Figure 1) to generate a sequence of spikes $S_{(t)}$ (which contains the channel numbers that generated each corresponding spike). From this sequence, we keep the N first spikes to perform a similarity comparison with each existing templates. This similarity is calculated by adding the weights of the model for each channel, modified by an inhibition factor that depends on the channel's rank in the sequence (Equ. 2) [11]. Finally, the template with the highest similarity to the actual model output is selected.

$$Similarity = \sum_{i=1}^{N} k_i \times I^{(rank\ of\ spike\ i)-1}, \tag{2}$$

where I stands for an inhibition factor (lower than or equal to one) and k_i for the corresponding weight of the template channel and threshold. As an example, with the same French digit "un" pronunciation (figure 2) and an inhibition factor of 0.9, the similarity (which is maximum) would be computed as follows : $20 \times 0.9^0 + 19 \times 0.9^1 + 18 \times 0.9^2 + ... + 2 \times 0.9^{18} + 1 \times 0.9^{19}$.

IV. EXPERIMENTS AND RESULTS

A. Speech Database

We performed a proof-of-concept test of speech recognition with our model, using an in house speech database made of 10 French digits spoken by 5 men and 4 women. Each speaker pronounced ten times the same digit (from 0 to 9). The speakers had a sequence of random numbers to read and the speech was recorded at 16 KHz using a headset.

B. Training and Recognition for our Prototype

For each digit, two reference models are used for the recognizer (one pronunciation for each sex). For each digit, for each sex and for each pronunciation inside the same sex group, a preliminary comparison between the same digits is performed. The comparison is made by computing the similarity measure (Equ. 2) between all pronunciations of the same digit inside the same sex group. The pronunciation with the highest similarity will be used as a reference model [3]. These reference models possess the highest similarity with the pronunciations in their respective group (digit, male or female). It should be noted that for this selection, the impact of a digit model on the pronunciations of the other digits (e.g., false recognition of other digit) is not taken into account.

Recognition has been performed on the ten pronunciations of each speaker. During recognition and for a given digit, seven speakers were not in the reference models.

C. Reference System

A conventional MFCC and Hidden Markov Model speech recognizer has been trained with the same training set than with the prototype [4]. The system uses hidden Markov models with 5 states for each digit (figure 3) and twelve cepstral coefficients for each time frame. The sliding window length is 32 ms with an overlap of 16 ms. For each state of the HMM, a mean and a variance are estimated during training. We use one Gaussian distribution of the observations (instead of the

[3]For example, the model of digit 1 for the male speakers is obtained: For $j = 1$ to 50 do (each digit 1 pronounced by a male – 50 pronunciations):

1) Compute the similarity with the other digits (49 similarities)
2) Compute the average

The pronunciation with the highest average similarity will be the model for digit "un" pronounced by the male speakers.

[4]The same reference pronunciation has been used for each digit.

conventional mixture of Gaussians) to reduce the number of parameters [5] to be estimated during training).

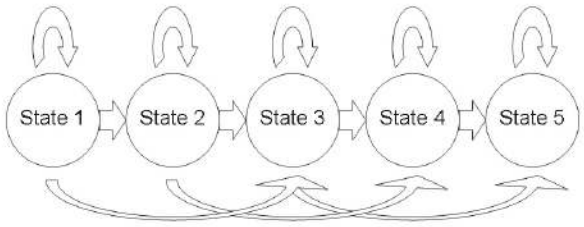

Fig. 3. Hidden Markov model.

D. Recognition Scores for Limited Data

TABLE II
RECOGNITION FOR EACH PRONUNCIATION OF THE TEN FRENCH DIGITS OBTAINED WITH OUR PROTOTYPE.

	Models										%
Number	1	2	3	4	5	6	7	8	9	0	64, 89
1 ("un")	84	1	4							1	93, 33
2 ("deux")		69	2	1				3	13	2	76, 67
3 ("trois")	10		58	18				1	1	2	64, 44
4 ("quatre")			22	68							75, 56
5 ("cinq")	11		1	2	42	21	6			7	46, 67
6 ("six")					1	68	13	2	1	5	75, 56
7 ("sept")		1	1	2	11	49	12	1	1	12	13, 33
8 ("huit")		1						68	16	5	75, 56
9 ("neuf")			4					9	54	23	60
0 ("zéro")					15	2	3	3	9	61	67, 78

Results from the prototype are reported on table II. Each row model gives the number of recognized pronunciations; each line is the pronounced digit to be recognized. The best score is obtained for digit "un" (usually short and difficult to be recognized with conventional speech recognizers). On the other hand, digits beginning with a similar fricative ("cinq", "six" and "sept") or plosives ("trois" and "quatre") are often confused. Since the neurons in our prototype fire only once, the emphasis is generally put on the first milliseconds of the signal. This could explain the observed confusions.

TABLE III
RECOGNITION FOR EACH PRONUNCIATION OF THE TEN FRENCH DIGITS OBTAINED WITH *HMM*S.

	Models										%
Number	1	2	3	4	5	6	7	8	9	0	51, 9
1 ("un")	15			1	20				28	16	16, 67
2 ("deux")		55			11			1	9	14	61, 11
3 ("trois")			42							48	46, 67
4 ("quatre")				33	32				10	15	36, 67
5 ("cinq")		2			81	1			1	5	90
6 ("six")					13	72		1		4	80
7 ("sept")					40	8	30			12	33, 33
8 ("huit")		31			5	8		5	38	3	5, 56
9 ("neuf")		4		10	19	3			44	10	48, 89
0 ("zéro")										90	100

[5]Approximately 130 parameters for each Markovian Model: Mean and Variance of a twelve dimensional vector for each state, 5 states and transitions between states.

The best score for the MFCC-HMM (table III) is obtained with digit number 5 (relatively long digit) and the worst with 1 (the best with our prototype) and 8. It is clear that digit model 8 is not correctly trained and that the HMM speech recognizer did not have enough data to correctly estimate the HMM parameters (approximately 130 parameters for each Markovian reference model).

E. Recognition with Bigger Training Set

Of course, with all 9 speakers in the training set, the HMM recognizer outperforms our Rank Order Coding prototype yielding an overall recognition rate close to 100%. For our prototype, the creation of reference models (with Maximum Likelihood, for example) is not trivial as it requires a new training rule to be created.

F. Discussion

When we use the same small training set (one pronunciation per digit and sex), the HMM is not able to converge during training and yields recognition rates of 50% (table III), while our speech recognizer prototype sits around 65% (table II).

For some short digits, the prototype yields interesting results (sometimes higher than with the MFCC-HMM recognizer). Even if the system is very crude in this initial implementation, interesting performance is obtained with a very limited training set. Moreover, good performance is observed for cases that are typically difficult for HMM recognizers, that is to say short words. While the HMM recognizer performance can be improved by increasing the training set, the performance of our prototype could benefit from various pre- or post-processing schemes that are currently under investigation (more realistic spike generation with integrate and fire neurons, learning rule, ...).

The resistance to noise of both presented systems were also found to be weak.

G. Recognition with Noisy Training Set

To evaluate the resistance to noise, tests at different SNR have been performed with both systems. The same clean limited training set selected in subsection IV-D is used and white noise is added to the recognition data set. In respect to robustness to noise with a limited training set, our prototype and the HMM present similar behavior. With an SNR of 20 dB, their averaged recognition rate are reduced by a factor of 50% in comparison to the results obtained with a clean limited training set. At this early stage, our prototype shows no real robustness to noise.

V. CONCLUSIONS

Conventional speech analysis and recognition techniques can yield good performance levels when correctly trained and when the test conditions match those of the training set. But for real-life situations, the designer has to train the system on huge databases that are very costly to implement. On the other hand, bio-inspired processing schemes can be unsupervised and generalize relatively well from limited data. They could

efficiently complement conventional speech processing and recognition techniques.

Due to the intrinsic spiking nature of neurons, suitable signal representations have to be found to adequately adapt the signal information to the neural networks. One important aspect of the bio-inspired prototype presented above is that it uses spikes trains. This places constraints on the type of coding that can be performed, and new signal representations had to be investigated to achieve computational efficiency and robustness to noise. Combined with conventional speech processing methods, this new approach could open up new research directions.

Improvements for our prototype include the use of a more complex neuron model. Indeed, for each neuron our prototype uses only the first spike of that neuron (equivalent to an infinite refractory period). Thus, emphasis is then generally put on the first milliseconds of the signal. With more complex neuron models, the produced spike trains could better characterize each word. The use of sparse representations [16], [17], [18], [19] could also be investigated for the same goal. In addition, the way our models are chosen and created is simple and fast, but it is far from optimized. Synaptic plasticity, as shown recently by Panchev and Wermter [20], can be used to perform recognition of sequences and could improve our recognition. Finally, A. Delorme et al. [21] have implemented a learning rule based on spike timing (spike timing dependent plasticity) in a visual framework that shows a great potential for the rank order coding. This learning rule should naturally be investigated for the creation of our models.

Acknowledgment

Many thanks to J. J. Rigoni, our COST277 collaborator, for stimulating discussions on speech recognition with spiking neurons, Hassan Ezzaidi for his help with the MFCC-HMM recognizer and Mr. Ferland for his proof-reading. Stéphane Loiselle has been partially supported by the 'Ministère de l'éducation du Québec' and by Université du Québec à Chicoutimi. Jean Rouat is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Simon Thorpe and Daniel Pressnitzer are supported by the "CNRS Integrative and Computational Neuroscience Initiative". This collaborative exchange between Québec and Europe has been partially supported by the MRST of Québec gvt and Université de Sherbrooke.

References

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 02 1989.

[2] "Speech communication journal", vol. 33, nb. 1–2, 2001.

[3] M. Hunt and C. Lefèbvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *ICASSP*, May 1989, pp. 262–265.

[4] M. Hunt, "Spectral signal processing for ASR," in *Proc. IEEE International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 12-15 1999.

[5] D. N. Zotkin, S. A. Shamma, P. Ru, R. Duraiswami, and L. S. Davis, "Pitch and timbre manipulations using cortical representation of sound," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2003, pp. 517–520.

[6] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition." *IEEE Tr. Speech and Audio Processing*, vol. 3, no. 1, pp. 115–132, 1994.

[7] S. Sandhu and O. Ghitza, "A comparative study of MEL cepstra and EIH for phone classification under adverse conditions," in *ICASSP 95*, vol. 1, 1995, pp. 409–412.

[8] P. Heil and D. Irvine, "First-spike timing of auditory-nerve fibers and comparison with auditory cortex," *J. Neurophysiol.*, vol. 78, pp. 2438–2454, 1997.

[9] M. Elhilali, J. Fritz, D. Klein, J. Simon, and S. Shamma, "Dynamics of precise spike timing in primary auditory cortex," *J Neurosci.*, vol. 24, pp. 1159–1172, 2004.

[10] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–533, june 1996.

[11] S. Thorpe, A. Delorme, and R. V. Rullen, "Spike-based strategies for rapid processing," *Neural Networks*, vol. 14, no. 6–7, pp. 715–725, 2001.

[12] R. VanRullen and S. J. Thorpe, "Surfing a spike wave down the ventral stream," *Vision Research*, vol. 42, no. 23, pp. 2593–2615, august 2002.

[13] L. Perrinet, "Comment déchiffrer le code impulsionnel de la vision? Étude du flux parallèle, asynchrone et épars dans le traitement visuel ultra-rapide." Ph.D. dissertation, Université Paul Sabatier, 2003.

[14] R. VanRullen, R. Guyonneau, and S. J. Thorpe, "Spike times make sense," *Trends in Neurosciences*, vol. 28, no. 1, p. 4, january 2005.

[15] R. Patterson, "Auditory filter shapes derived with noise stimuli," *JASA*, vol. 59, no. 3, pp. 640–654, 1976.

[16] B. A. Olshausen, *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, 2002, ch. 13: Sparse Codes and Spikes, pp. 257–271.

[17] B. A. Olshausen and K. J. Millman, "Learning sparse codes with a mixture-of-gaussians prior," in *NIPS*, vol. 12, 1999, pp. 841–847.

[18] C. Feldbauer and G. Kubin, "How sparse can we make the auditory representation," in *ICSLP 2004*, October 2004.

[19] M. S. Lewicki, *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, 2002, ch. 12: Efficient Coding of Time-Varying Signals Using a Spiking Population Code, pp. 243–255.

[20] C. Panchev and S. Wermster, "Spike-timing-dependent synaptic plasticity: from single spikes to spike trains," in *Computational Neuroscience Meeting*. Springer–Verlag, july 2003, pp. 494–506.

[21] A. Delorme, L. Perrinet, and S. Thorpe, "Networks of integrate-and-fire neurons using rank order coding b: Spike timing dependent plasticity and emergence of orientation selectivity," *Neurocomputing*, vol. 38–40, pp. 539–545, 2001.