

2

A Prototype Speech Recognizer based on Associative Learning and Nonlinear Speech Analysis

Jean Rouat and Miguel Garcia
Université du Québec à Chicoutimi

The grouping of sounds has been shown to be partially based on amplitude modulation (AM) characteristics (Bregman et al., 1985), suggesting that AM information observed in auditory nerve fibers could be used by the auditory system to segregate speech from background noise. This chapter proposes a speech recognizer prototype that relies on patterns of modulation observed in auditory fibers influenced by a summation of close harmonics. The recognition task is performed by a modified Dystal (DYNamically STable Associative Learning) neural network (Alkon et al., 1990) (Blackwell et al., 1992). Preliminary results indicate that the approach might be efficient and powerful. Further experiments have still to be done in order to evaluate the approach. Experiments on continuous speech in noisy environment are planned.

2.1 INTRODUCTION

The analysis and the recognition of speech spoken in noisy environment is a difficult and crucial task. Most of the speech recognizers alleviate the difficulties of this task by training on noisy data, assuming that the statistical properties of the noise will remain unchanged between the learning and the recognition phase. Other techniques assume that the speech source and the interference noise sources are spatially different which is not necessarily so. In summary, the most effective techniques are generally useful for specific conditions.

On the other hand, perceptive analysis and auditory models enhance the discriminant information from non stationary noise and are supposed to yield

A Prototype Speech Recognizer

good performance in adverse conditions. However, their complexity and the difficulty of exploiting the dynamic output information with standard pattern recognition algorithms restrict their integration in speech recognizers.

According to Bregman (Bregman, 1984), the phasic analysis performed by the auditory system, in conjunction with the tonal analysis, is adapted to the perception of speech in adverse environment. The spectral integration (or grouping of sounds) was shown to be partially based on common amplitude modulation characteristics (Bregman et al., 1985).

Furthermore, research work on automatic demodulation of speech can be motivated by the fact that the human brain has neural cells specialized in AM and frequency modulation (FM) detection (Gardner & Wilson, 1979) (Tansley & Suffield, 1983). Robles, Ruggero and Rich (Robles et al., 1991) observed distortion products on chinchilla basilar membrane and they suggested that the living basilar membrane is a nonlinear system. Thus, the perception of distortion products could be due to the basilar membrane response and not only to neural postprocessing. Moreover, simple nonlinear operators can enhance the AM or FM information in a signal (Maragos et al., 1992) (Rouat, 1993) and can be used to process the output of a cochlea filterbank in order to obtain AM information characteristics of speech signal and segregate it from background noise (Rouat et al., 1992).

2.2 MODULATION IN THE AUDITORY SYSTEM

The modulation information is one of the main cues extracted by the auditory system. Schreiner and Langner (Langner & Schreiner, 1988) (Schreiner & Langner, 1988) showed that the inferior colliculus of the cat contains a highly systematic topographic representation of AM parameters and maps showing 'best modulation frequency' have been determined.

The poor spectral resolution of the cochlea can be an advantage when the speech signal is harmonic, as more than one harmonic of the signal can fall into the same channel producing an amplitude modulated signal with a modulation frequency equal to (or a multiple of) the fundamental frequency. Therefore, the fundamental frequency F_0 (or a multiple of F_0) can be encoded in the temporal discharge patterns of auditory nerve fibers. The characteristic frequencies of these nerve fibers can be very different from the fundamental frequency F_0 (or a multiple of F_0).

As the bandwidth of the nerve fibers vary significantly (even for fibers tuned to the same frequency) and since the bandwidth may be broad, at least some of the auditory nerve fibers might be able to encode the envelopes relevant to speech signals. Langner (Langner, 1992) showed how such periodicity coding is related to modulation information and analyzed the role of the "On"

A Prototype Speech Recognizer

and "Chopper" neurones (in the cochlear nucleus) as preprocessors for enhancing the AM information coming from the auditory nerve fibers.

2.3 APPLICATION TO VOICED SPEECH

Many studies have concentrated on the coding of vowels (Delgutte, 1980) (Delgutte & Kiang, 1984) in the auditory nerve. For the nerve fibers whose characteristic frequency (CF) is close to a formant frequency, a phase-coupling to the formant frequency or to an adjacent harmonic is observed with little or no envelope modulation as the discharge pattern of the fiber is dominated by a single large harmonic component. Other fibers may show modulations corresponding to harmonic interactions. Therefore, the auditory system is able to track simultaneously formants and pitch by relying on phase-coupling of fibers whose CF is close to the formant ('spectral analysis') and by relying on patterns of modulation for fibers influenced by a summation of stimulus harmonics (Delgutte, 1980) (Miller & Sachs, 1984).

Most speech recognizers are based on short-term analyses that assume that the speech signal in the analysis window (typical duration of 10 to 20 ms) is stationary. As a consequence, those systems can not exploit some of the instantaneous characteristics of speech. In fact, a short-term Fourier (or linear predictive coding) analysis estimates the averaged values of harmonics on a short speech segment and can not extract accurately the patterns of modulation created by close harmonics. As these patterns seem to characterize what has been pronounced, and because they occur at glottal explosions, short-term analysis fail to exploit information that might be very useful for recognition purposes. On the other hand, most perceptive or auditory based models require a large number of inner hair cell models or cochlea filters in order to obtain a sufficiently accurate estimation of the frequency distribution of speech. To our knowledge, not much work has been done to exploit the patterns of modulation observed in the auditory nerve for formants interaction estimation and for pitch estimation.

We assume here that some of the cochlear filters and auditory nerve fibers have bandwidths broad enough to encode the envelope of the modulation produced by interacting harmonics close to formants (F1 and F2 in /a/, F2 and F3 in /i/, F3 and F4 in /i/, etc.). Therefore, the analysis does not require many cochlear filters (24) and yields an estimation of $F2 - F1$, $F3 - F2$ or $F4 - F3$ and of the missing fundamental by extracting the modulation pattern at the output of cochlear filters, when possible. Furthermore, there is no need to assume any stationary signal.

This system exploits the modulation information in order to perform the recognition of voiced speech segments based on patterns of modulation using a Dystal neural network (Alkon et al., 1990). We are interested in the patterns of

A Prototype Speech Recognizer

modulation produced by 'interacting' harmonics (beats of harmonics) in medium and high frequency channels of a bank of cochlear filters in order to classify vowels and voiced sounds. The Dystal network uses the 3D speech representation delivered by the analysis to classify images that are characteristics of voiced speech.

2.4 THE ANALYSIS

The filterbank is comprised of a bank of 24 filters centered on 330Hz to 4700Hz (Moore & Glasberg, 1983)(Patterson, 1976). The output of each filter is a bandpass signal with a narrow-band spectrum centered around f_i where f_i is the central frequency (CF) of channel i. The output signal $s_i(t)$ from channel i can be considered to be modulated in amplitude and phase with a carrier frequency of f_i .

$$s_i(t) = A_i(t)\cos[\omega_i t + \phi_i(t)] \quad \text{EQ. 2.1}$$

$A_i(t)$ is the modulating amplitude and $\phi_i(t)$ is the modulating phase.

In this chapter we use a finite impulse response time digital Hilbert transformer (Rabiner & Schafer, 1974) to extract the envelope of $s_i(t)$. Thus

$$A_i(t) = \sqrt{s_i(t)^2 + s_i(t)_q^2} \quad \text{EQ. 2.2}$$

where $s_i(t)_q$ is the Hilbert transform of $s_i(t)$. Then, an image representation is obtained by plotting the product $A_i(t) \cdot A_i(t)'$ versus time and central frequency. $A_i(t)'$ is the time derivative of $A_i(t)$. The x axis is the time and the y axis is expressed in hertz according to the ERB (equivalent rectangular bandwidth) scale (Patterson, 1976). The image color is the $A_i(t) \cdot A_i(t)'$ variable.

A Prototype Speech Recognizer

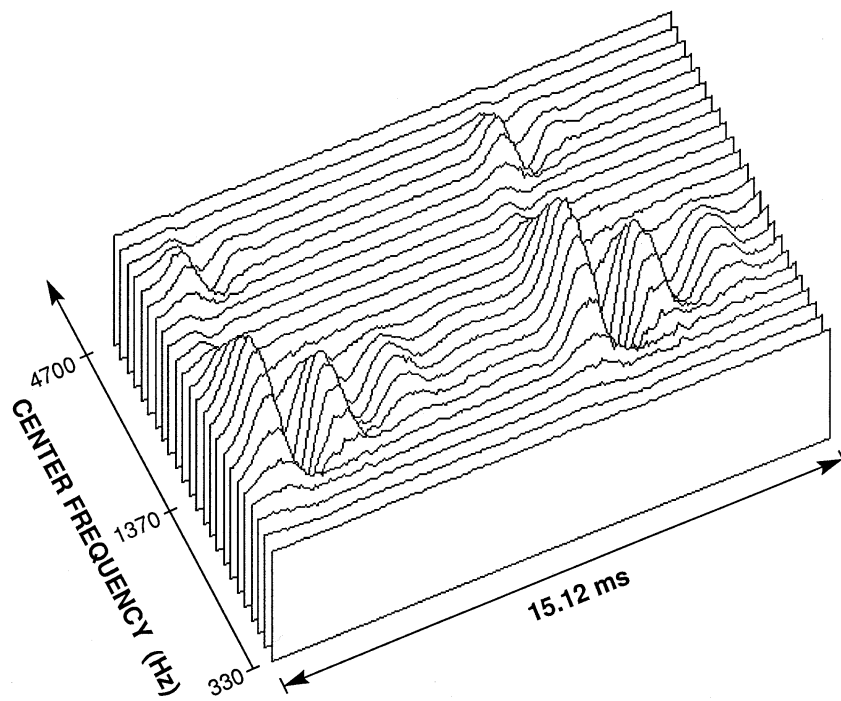


FIG. 2.1. /a/ speech segment, male speaker.

Figure 2.1 shows two pitch periods taken from a French vowel /a/ pronounced by a male speaker. The modulation occurs during glottal explosion. It is due to the interaction of harmonics close to F_1 and F_2 (in channels 8 to 13) and to F_3 and F_4 in channels 19 to 21. The period of the modulation is a multiple of T_0 and is approximately equal to $1/(F_2 - F_1)$ (or to $1/(F_4 - F_3)$) and is independent of noise power. Figure 2.2 is a three pitch periods of the French vowel / ϵ /. It has been pronounced by a female speaker. The modulation pattern is essentially due to F_2 and F_3 (in channels 17 to 19).

A Prototype Speech Recognizer

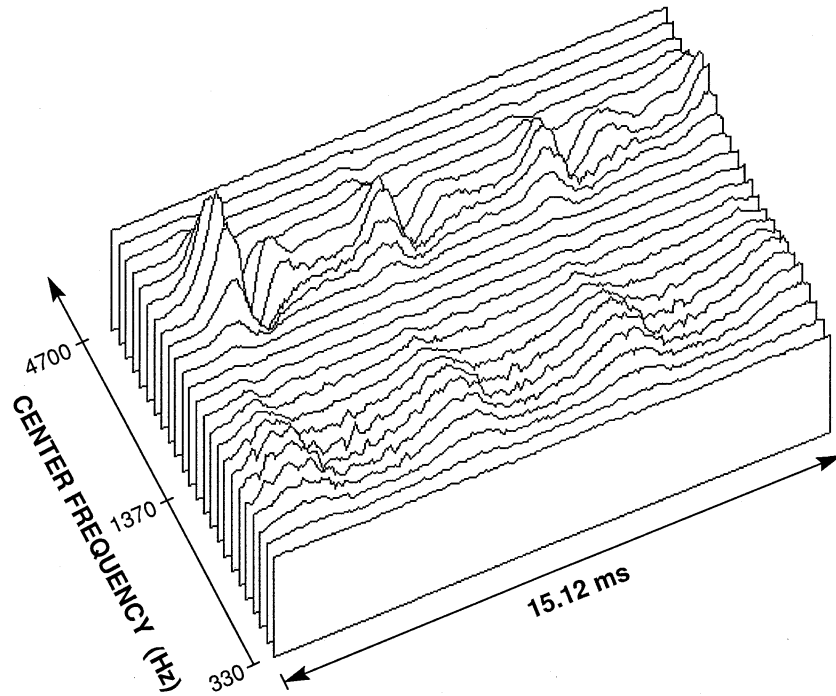


FIG. 2.2. /ε/ speech segment, female speaker.

2.5 THE RECOGNIZER

2.5.1 The Dystal Neural Network

The recognizer is based on an artificial network derived from a marine snail and the hippocampus of a rabbit. The Dystal network was developed by Alkon et al. (Alkon et al., 1990). Experiments using Dystal were already reported on handwritten character recognition (Blackwell et al., 1992).

Alkon et al. (Alkon et al., 1990) proposed a network that associatively learns correlations and anticorrelations between time events occurring in presynaptic neurons. Those neurons synapse on the same element of a common postsynaptic neuron. They proposed a learning rule that modifies the cellular excitability at dendritic patches. These synaptic patches were postulated to be

A Prototype Speech Recognizer

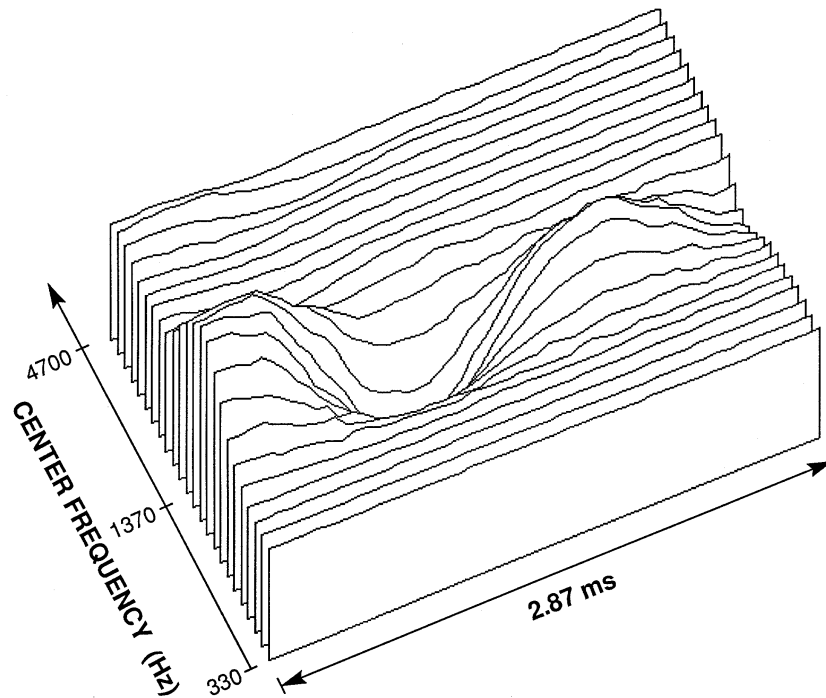


FIG. 2.3. /a/ reference pattern, male speaker.

formed on branches of the dendritic tree of vertebrate neurons. In Dystal, weights are associated to patches rather than to incoming connection. After learning, each patch characterizes a pattern of activity on the input neurons. A Dystal network has two separate input pathways. A first pathway is the conditioned stimulus (CS) input and the second is the unconditioned stimulus (UCS). The CS input neurons are the receptive field of a Dystal neurone. The UCS inputs are used during learning and are associated to specific patterns of CS inputs.

2.5.2 The learning and the pattern recognizer architecture

The basic architecture. In training mode, Dystal learns patches corresponding to selected images (output of the multichannel analysis). The selected images are reference patterns and characterize the $A_i(t) \cdot A_i(t)'$ product for specific vowels. For example, Figures 2.3, 2.4 and 2.5 are reference patterns for the French vowels /a/, /i/ and /ε/.

A Prototype Speech Recognizer

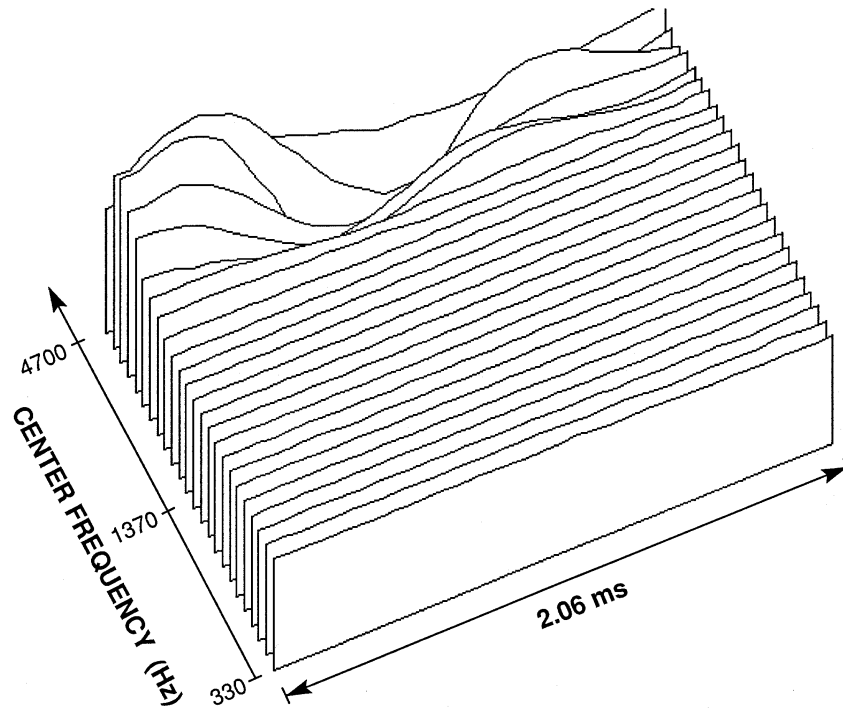


FIG. 2.4. /i/ reference pattern, male speaker.

For each patch that is being learned, an unconditioned stimulus (UCS) (a pattern name) is presented simultaneously with a conditioned stimulus (CS) (a reference pattern). For example, the pattern of Fig. 2.3 is presented on the CS neurons input in association with the symbol /a/ as UCS input. This allows Dystal to associate the CS input pattern with the vowel /a/. For each CS image input, Dystal computes a similarity value with each of the other patterns that are already associated with the same UCS input. For example, when the UCS is /a/, Dystal computes similarities between the actual CS input and the other “a-patterns” stored in memory. If all the similarities are beneath a presettled threshold, a new pattern is created by storing the CS neurons values in patches. The patches comprise the reference pattern images and the corresponding UCS entries. If one of the similarities is above the threshold, no pattern is stored. The Dystal learning phase we use is described in (Blackwell et al., 1992).

During the recognition phase the spectro-temporal image of the spoken speech is computed. This image can be obtained in real time as there is no constraint on the duration of speech. The system generates the image for short speech segments as well as for long fluent speech sentences. A variable length window is shifted along the spectro-temporal image. For each position of the

A Prototype Speech Recognizer

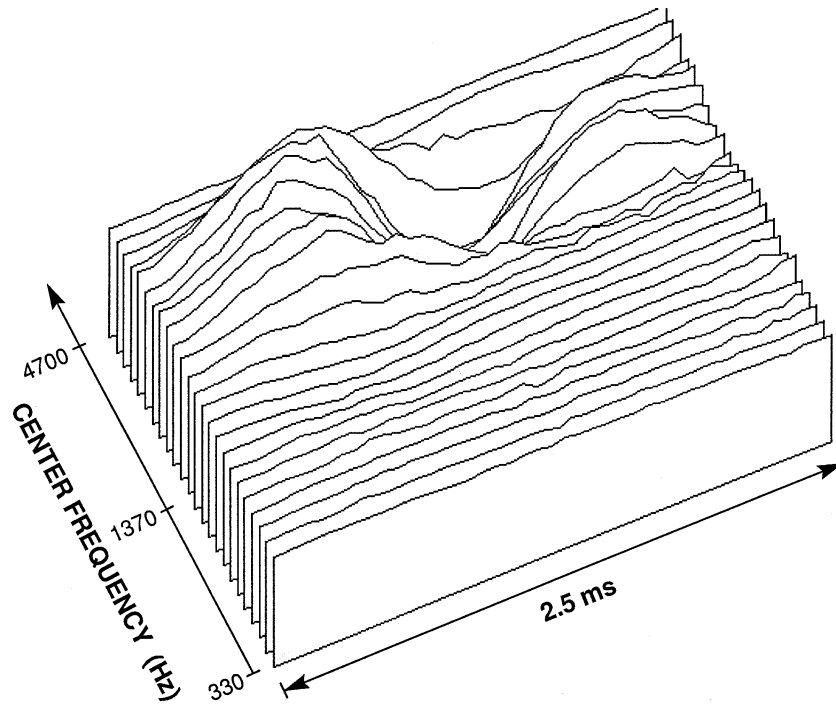


FIG. 2.5. / ϵ / reference pattern, female speaker.

window a recognition is performed. The UCS entry of the closest reference pattern is given as output if there is a strong similarity between the reference patch and the data presented to the CS inputs. More precisely, the CS input neurons receive the data of the window with no associated UCS. The data of the window are compared with the reference patterns stored in Dystal memory. A similarity measure is computed for each class of vowels. The UCS stimulus associated with the greatest similarity is presented as output when that similarity is sufficiently high.

The two layers architecture. The two layers architecture includes the basic system as layer one. Learning and recognition are performed in the same manner and with the same conditions. In opposition to the previous architecture, no recognition decision is taken by the first layer. The output of layer one is a vector of similarity measures. The dimension of the vector is equal to the number of classes the system has to recognize. Each component of the vector is the similarity of a specific class. A vector of similarity measures is generated for each window. Consequently, the output of layer one is the time evolution of

A Prototype Speech Recognizer

the vector of similarities. The prototype version generates a vector of similarities every 1/16 ms (i.e. the window shift is equal to 1/16 ms).

Layer two uses a fixed length window (10 ms) that is shifted (5 ms) across the image obtained by representing the time evolution of the output of layer one. After each window shift, the local maximum of the image in the window is located. A new segment that begins at the maximum location and ends 2 ms later is extracted. Similarly, for each 10 ms portion of image, a 2 ms segment is extracted following the local maximum. The segments are then concatenated. The concatenation creates a static image of similarity measures that is independent of the fundamental frequency.

A typical reference pattern for layer two consists of 10 sections of 2 ms images. The learning for layer two is exactly the same as for layer one. The difference resides in the CS inputs, that are the output images of layer one. The UCS are also the class names.

2.6 PRELIMINARY EXPERIMENTS AND RESULTS

A preliminary experiment has been conducted on four vowel classes: /a/, /i/, /y/ and / ϵ /. Six speakers were used for training (3 males and 3 females). A different male speaker was used for recognition. The reported results were obtained with the basic architecture.

2.6.1 Learning speakers

For each class, the reference speakers were randomly chosen. Therefore, the classes are not characterized by the same speakers. The number of reference patterns was randomly chosen. The /a/ and /i/ vowels are respectively represented by 3 reference patterns taken from 2 males and 1 female. The /y/ and / ϵ / vowels are also characterized by 3 reference patterns taken from 2 other females and 1 other male. The basic architecture of the system was tested with 1 male speaker in a speaker-independent mode. In the first experiment, the speaker pronounced the isolated vowels /a/, /i/, /y/ and / ϵ /, and the system had to recognize them. In the second experiment the speaker pronounced the isolated letters of the French alphabet.

The speech was sampled at 16000 kHz prior to be filtered by the cochlear filterbank. A speech image representation was then generated for each letter. A variable length window (2 ms to 3 ms) was placed on the spectro-temporal image. The window was shifted every 1/16 ms. For each pattern stored in Dystal, a similarity measure was computed once the window was positioned. When computing a similarity measure, the window length was automatically adapted to the length of the reference pattern under comparison.

2.6.2 Results

The recognition of /a/, /i/, /y/ and / \mathcal{E} / in a speaker-independent mode yielded a recognition rate of 100%. The preliminary results of the second experiment are reported below:

- The vowel segments of letters a and k were recognized as /a/.
- The vowel segments of letters f, l, m, n, r and s were recognized as / \mathcal{E} /.
- The vowel segments of letters i, j, x and y were recognized as /i/.
- The vowel segments of letters q and u were recognized as /y/.
- The vowel segments of letters b, c and d were recognized simultaneously as /i/ and /y/.
- The vowel / Φ / from letter e was recognized simultaneously as /y/ and / \mathcal{E} /.
- The unvoiced segments of letters f, c and s were recognized as /i/.

2.7 DISCUSSION

The preliminary results confirm that the modulation information observed during glottal explosion might be useful for classification and recognition of speech. The task presented in this paper is relatively easy and could be performed very well with standard speech recognizers (dynamic time warping, hidden Markov models, or neural networks). In fact, vowels are relatively stable and can be characterized with the distribution of averaged spectral energy. The prototype system we propose performs a very different speech analysis and exploits different information. The spectral resolution of the analysis we use is very poor. Consequently, the spectral distribution cannot be used (except for distinguishing /a/ from the 3 other vowels). But the temporal resolution is very good, and the vowels /i/, /y/ and / \mathcal{E} / are recognized based on the time information. This information resides in the period of the envelope of the modulated signal. It easily can be extracted and seems to be more robust to noise than the energy distribution. The training of the system is fast, once the reference patterns have been defined.

The modulation information seems to be reliable and robust to noise (Rouat et al., 1992). In comparison with standard speech recognizers (dynamic time warping, hidden Markov models, or neural networks), the system is very different. The robustness of contemporary noisy speech recognition systems is mainly due to signal processing techniques at the level of the input signal or to specific strategies embedded in the recognition process. Signal processing techniques are usually used to clean the signal and remove noise. The recognition strategies include training and learning on noisy data, estimation of the noise statistical properties, inclusion of model of noises, and so on. The robustness of the proposed system is due to the acoustical cues obtained via the modulation analysis. Therefore, the recognition process can be more simple.

A Prototype Speech Recognizer

Some vowels or speech classes that were never learned by the network seem to be characterized by simultaneous responses in two or more Dystal output neurons. Depending on what was pronounced, the neurons that are activated are different. Some classes that were never learned might be characterized by a strong response appearing in different output neurons. In that case, layer 2 might help to improve the potential of the approach.

2.8 CONCLUSION

We have proposed a new speech analysis, and we designed a prototype speech recognizer to evaluate the speech demodulation approach. We exploit modulation cues that are characteristics of speech and that can be used for noisy environments and auditory scene analysis. The system is based on events that are present in speech and that cannot be taken into account by the majority of contemporary speech systems.

Further experiments with many speakers and with a large vocabulary should be performed in order to evaluate the potential of that approach to speech recognition and auditory scene analysis. Furthermore, evaluations in various noisy environments have to be performed.

This chapter reports results on the exploitation of modulation information for voiced speech. More experiments have to be done in order to evaluate the pertinence of a similar approach to unvoiced speech. For unvoiced speech, the time evolution of the envelope $A_i(t)$ is a meaningful cue. It is not used by most speech systems because of the difficulty of obtaining the envelope on short segments. The processing of unvoiced speech by our system will be studied in the near future.

The prototype system has been implemented in order to study the modulation in speech and has not been yet optimized in terms of CPU time. Improvements can be done by increasing the window shift and reducing the order of the filters.

ACKNOWLEDGMENTS

This work has been supported by the National Sciences and Engineering Research Council of Canada, by the Fonds pour la Formation des Chercheurs et l'Aide à la recherche du Québec, by the Canadian Microelectronics Corporation and by the Fondation from Université du Québec à Chicoutimi. Many thanks to Daniel Morissette for his programming work.

REFERENCES

- Alkon, D. L., Blackwell, K. T., Barbour, G. S., Rigler, A. K., & Vogl, T. P. (1990). Pattern-recognition by an artificial network derived from biologic neuronal systems. *Biological Cybernetics*, 62, 363-376.
- Bregman, A. S. (1984). Auditory scene analysis. *Proceedings of the seventh international conference on pattern recognition*, pp. 168-175. Silver Spring, MD: IEEE Computer Society Press.
- Bregman, A. S., Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Journal of Perception and Psychophysics*, 37, 483-493.
- Blackwell, K. T., Vogl, T. P., Hymans, S. D., Barbour, G. S., & Alkon, D. L. (1992). A new approach to hand-written character recognition. *Pattern Recognition*, 25 (6), 655-666.
- Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. *Journal of the Acoustical Society of America*, 68, 843-857.
- Delgutte, B. & Kiang, N. Y. (1984). Speech coding in the auditory nerve: Vowels in background noise. *Journal of the Acoustical Society of America*, 75, 908-918.
- Gardner, R. B. & Wilson, J. P. (1979). Evidence for direction-specific channels in the processing of frequency modulation. *Journal of the Acoustical Society of America*, 66, 704-709.
- Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research*, 60 (2), 115-142.
- Langner, G. & Schreiner, C. E. (1988). Periodicity coding in the inferior colliculus of the cat. Neuronal mechanisms. *Journal of Neurophysiology*, 60 (6), 1799-1822.
- Maragos, P., Quatieri, T. F., & Kaiser, J. F. (1992). On separating amplitude from frequency modulations using energy operators. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2.1-2.4.
- Miller, M. I. & Sachs, M. B. (1984). Representation of voice pitch in discharge patterns of auditory nerve fibers. *Hearing Research*, 14, 257-279.

A Prototype Speech Recognizer

- Moore, B. & Glasberg, B. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74, 750-753.
- Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 59 (3), 640-654.
- Rabiner, L. R. & Schafer, R. W. (1974). On the Behavior of Minimax FIR Digital Hilbert Transformers. *The Bell Systems Technical Journal*, 53 (2), 363-390.
- Robles, L., Ruggero, M. A., & Rich, N. C. (1991). Two-tone distortion in the basilar membrane of the cochlea. *Nature*, 349 (6308), 413-414.
- Rouat, J., Lemieux, S., & Migneault, A. (1992). A spectro temporal analysis of speech based on nonlinear operators. *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, 1629-1632 edited by J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodege, & G. E. Wiebe: the university of Alberta, Edmonton, Canada.
- Rouat, J. (1993). Nonlinear operators for speech analysis. *Visual Representations of Speech Signals*, edited by M. Cooke, S. Beet and M. Crawford. J Wiley. pp. 335-340.
- Schreiner, C. E. & Langner, G. (1988). Periodicity coding in the inferior colliculus of the cat. Topographical organization. *Journal of Neurophysiology*, 60 (6), 1823-1840.
- Tansley, B. W. & Suffield, J. B. (1983). Time course of adaptation and recovery of channels selectively sensitive to frequency and amplitude modulation. *Journal of the Acoustical Society of America*, 74, 765-775.