

Variable Frame Rate Hierarchical Analysis for Robust Speech Recognition

Jean Rouat*, Stéphane Loiseau* and Stéphane Molotchnikoff⁺
NECOTIS, GEGI, *Univ. de Sherbrooke, ⁺ Univ. de Montréal, QC, CANADA

Abstract—A new bio-inspired speech analysis system that extracts acoustical speech events is proposed and used in the design of a variable frame rate (VFR) speech recognizer. The same speech recognizer (Hidden Markov Model -HMM- and Mel Frequency Cepstrum Coefficients -MFCC-) has been used with the proposed VFR analysis and conventional fixed frame rate (FFR) approach. In comparison with other VFR recognizers, the hierarchical features in the proposed system have the potential to serve as classification parameters of a complete bio-inspired speech recognition system. Also, no voice activity detection is required and there are no hard decisions to be taken by the system. Events are used to label and identify the moments at which the acoustical properties of speech are stable or changing. These events are markers on which an analysis window can be positioned to perform the recognition. Inspired by our knowledge of the auditory and visual systems, hierarchical complex features like transients and energy orientation are used. Training has been done on clean speech and recognition on noisy (from 20dB to -10dB Signal to Noise Ratios -SNR) or reverberated speech by using the *TI 46-word* database corrupted with 4 noises taken from the *Aurora 2* data. In comparison with a FFR recognizer, our VFR system yields more than 50% increase in recognition rates for a speaker independent isolated word recognition task when SNRs are between 0 and 20 dB.

I. INTRODUCTION

A. Requirements for Robot Speech Recognition

Training in noisy environment where the noise is included in the statistics of the models has been a popular practice for almost two decades [1]. But robots must be robust to unseen environmental conditions and original methods have to be developed so that robots obtain reasonable speech recognition performance with limited training.

Current robot speech recognizers use an overlapping fixed length sliding window. Therefore, they cannot exploit the short-term non stationary property of speech. Speech is redundant and few short segments of signal are sufficient to identify what has been said. Reasonable recognition scores are obtained even if the amount of training data is limited when the analysis focuses on transients or onsets (see [2], [3] for a simple preliminary study).

This work is to be related to previous Variable Frame Rate (VFR) analyzes that are proposed in the literature [4], [5], [6], [7]. State of the art VFR estimate variation or stability of predefined parameters like energies, zero crossing rates, pitch frequencies, wavelet coefficients. Decisions (keep or reject the current frame) are based on hard thresholding. Our proposal explores the potential of a new hierarchical speech analysis that focuses on signal frames that carry "meaningful" and "robust" features whether they are stable

or transients. When taking into account the intrinsic non-stationarity of speech, it is possible to exploit short-time intervals in which instantaneous SNR are sufficiently high to generate reliable features.

B. Acoustical Event Detection

We propose a system that extracts acoustical events from speech. They correspond to onsets and to moments at which our hierarchical features are stable, ascending or descending. These events are markers on which an analysis window may be positioned to perform speech recognition. Results show that the approach greatly improves the recognition rates of noisy or reverberated speech.

II. DESCRIPTION OF THE HIERARCHICAL ANALYZER

A. Speech Analysis Module

The signal is first filtered by a 120-channel finite impulse response cochlear filterbank¹ distributed according to the Bark scale [14]. Then, the envelopes of the filter outputs are computed as the modulus of the Hilbert transform and are compressed with a logarithm function. This spatio-temporal representation is our first level simple features (Fig 1b) and is noted as $S(t, c)$ where t is time and c is the cochlear channel index.

1) *First Level Complex Features Extraction*: The first level complex features $Co(t, c)$ (Fig. 1c) are obtained by projecting the first simple features on 5 weighted temporal wavelets and by taking the maximum of the 5 scalar products. These five temporal wavelets enhance the temporal modulations of the simple first-level features (Fig. 2).

The temporal basis ($b_{f_b}(t)$) (Eq. 1) are damped sine waves with frequencies covering the range of spatio-temporal receptive fields found in the primate auditory system (also corresponding to the averaged syllable and word rates of speech).

$$a_{f_b}(t) = e^{-t} \sin(2\pi f_b t), \text{ with } f_b = (2, 4, 8, 16, 32) \text{ Hz} \quad (1)$$
$$b_{f_b}(t) = a_{f_b}(t) - \text{mean}(a_{f_b}(t))$$

$$Co(t, c) = \max_{f_b=2,4,\dots,32} [S(t, c) b_{f_b}(t)] \quad (2)$$

These complex first level features, $Co(t, c)$ are in some sense the maximum changes of the energy of the simple first level feature representation for scales 2, 4, 8, 16 and 32 Hz. As the 5 basis have different durations, a different gain is associated to each base (not illustrated in figure 2).

¹implemented by S.Gagné, Y.Liu and J.Rouat, inspired by the works of R.Patterson [13]

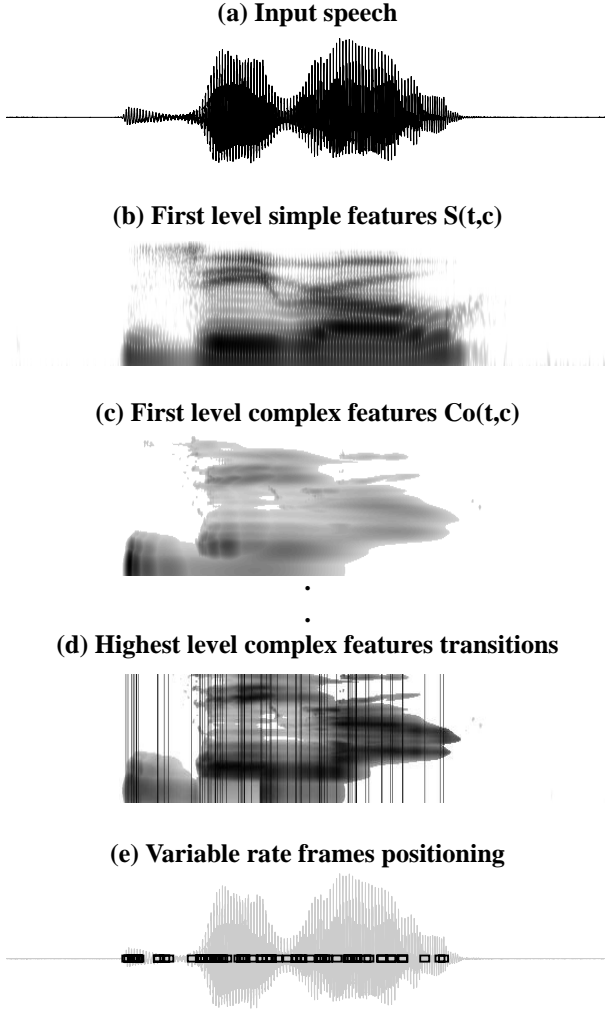


Fig. 1. Hierarchical features. (a) Raw speech input (French digit /z//e//r//o/), (b) first level simple features (cochleogram), (c) first level complex features (pooled coefficients of five transient wavelets), (d) timing markers obtained from the highest features and illustrated on the first level complex features, (e) positions of frames to be used by the speech recognizer. No segmentation or voice activity detection are needed. The system finds reliable signal frames.

The highest gain is given to the shortest base and the smallest gain to the longest base.

2) *First Level Complex features and Representation Generation*: In next step, the complex first level features $Co(t, c)$ are combined with the simple features $S(t, c)$. Before combination, complex features $Co(t, c)$ are normalized by the greatest $co_c(t)$ ($c = 1, \dots, 120$), at each instant t . c is the index of the cochlear channel and $co_c(t)$ is a coordinate of the complex feature vector with $Co^T(t, c) = [co_1(t), \dots, co_i(t), \dots, co_{120}(t)]$. T is the transposition operator. At each instant t , the new normalized complex features (denoted as $Co_N(t, c)$) are computed based on the original $Co(t, c)$.

$$Co_N^T(t, c) = \frac{1}{\max_{c=1, \dots, 120} (co_c(t))} [co_1(t), \dots, co_i(t), \dots, co_{120}(t)]$$

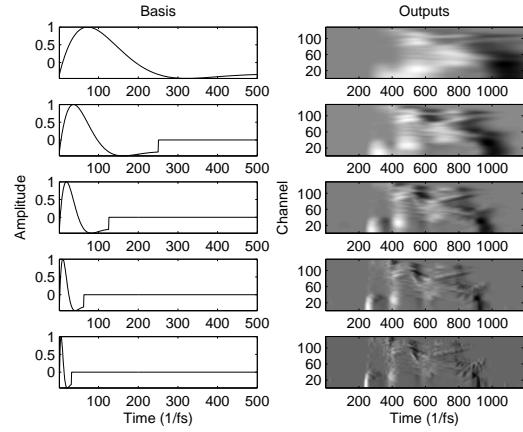


Fig. 2. Temporal basis $b_{f_b}(t)$ (left) and their outputs (right) for the same utterance as in Fig. 1. From top to bottom, the temporal basis center frequency is 2, 4, 8, 16 and 32 Hz.

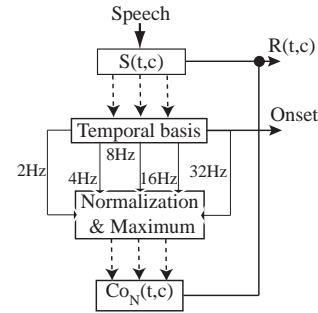


Fig. 3. First level features generator: ONSET information and spatio-temporal normalized representation $R(t, c)$

Then, the final representation is the product through time of $Co_N(t, c)$ and $S(t, c)$ yielding

$$R^T(t, c) = [co_{N,1}(t)s_1(t), \dots, co_{N,120}(t)s_{120}(t)]$$

The first level feature generator yields a normalized bi-dimensional representation $R(t, c)$ and the Onset information (assumed here as being correctly identified by the output projection of $S(t, c)$ on the 32 Hz base).

B. Second Level Feature Extraction

The higher we go in the hierarchy of the auditory system, the higher the time constants become (longer time integration). We now look at second-level features that are based on the time course evolution of orientations (stationary, ascending or descending energies, changes in the orientation) of the first level complex features $R(t, c)$. When speech is voiced, these features can be assimilated to pseudo-formants.

1) *2nd level simple features*: For each time instant t , a Gaussian window $G_{\bar{c}}(c) = e^{-\frac{1}{2} \frac{(c-\bar{c})^2}{32}}$ with a length of 29 channels is shifted along the cochlear channel axis (\bar{c} and c indexes) and correlated with $R(t, c)$. Then the cochlear center frequency of the Gaussian center \bar{c} that maximizes the correlation (implemented as a scalar product because of the symmetry of the Gaussian window) is used as a pseudo

first formant F_1 estimator.

$$F_{1,idx} = \underset{\substack{\bar{c}=1,\dots,120 \\ \text{with } c=1,\dots,120}}{\operatorname{argmax}} [G_{\bar{c}}(c)R(t, c)] \quad (3)$$

Once the position of the pseudo-formant is known ($F_{1,idx}$), the value of $R(t, F_{1,idx})$ is replaced by the sum of the values in $R(t, c)$ from $c = 1$ to $F_{1,idx} + 14$. The $R(t, c)$ for $c = 1$ to $F_{1,idx} + 14$ (except for channel $F_{1,idx}$) are erased.

$$R(t, F_{1,idx}) = \sum_{c=1}^{F_{1,idx}+14} R(t, c) \quad (4)$$

$$R(t, c) = 0; \text{ for } c = 1, \dots, F_{1,idx}-1, F_{1,idx}+1, \dots, F_{1,idx} + 14 \quad (5)$$

This operation is equivalent in transferring the energies from channels 1 to $F_{1,idx} + 14$ to channel $F_{1,idx}$. A narrow Gaussian basis gives a good resolution in frequency, but has to be broad enough so as not to detect multiple peaks in the same frequency area.

Then a second peak position is estimated (position of the pseudo F'_2 : see Chistovich [15] and Fant [16], [17]). At each instant t , the center of mass of the first level complex features are estimated based on the updated values of $R(t, c)$ with $c = F_{1,idx}+15$ to 106. The channel frequency of the center of mass is then attributed to F'_2 and the index of that cochlear channel is noted as $c = F_{2,idx}$. As for F_1 and for a fixed instant t , the sum of the $R(t, c)$ with $c = F_{1,idx}+15, \dots, 106$ is attributed to $R(t, F_{2,idx})$.

$$R(t, F_{2,idx}) = \sum_{c=F_{1,idx}+15}^{106} R(t, c) \quad (6)$$

In some situations, there exists a strong gap between F_1 and F'_2 (f.e. phoneme /i/). Once F'_2 is estimated, a minimum is searched between F_1 and F'_2 . If there exists one, it is kept, otherwise it is ignored. Fig. 4 illustrates the intermediate results for a French /i/ by a female speaker.

2) *Second Level Complex features*: Ascending and descending tendencies of the simple features F_1 and F'_2 are now identified and characterized. An ascending tendency is a trajectory short-term increase in frequency. A short-term decrease in frequency is qualified as a descending tendency. In the case where change in frequency trajectory is minimal, it is considered as invariable. Each trajectory F_1 or F'_2 is filtered out with seven Gabor filters (Fig. 5) to generate 3 kinds of second level orientation complex features [18]. For each trajectory F_1 and F'_2 and at each instant t , three complex feature coefficients are obtained: one coefficient for each feature (stable, ascending and descending). The coefficients for the ascending and descending complex features are obtained by max-pooling (keep the maximum) over the output of each 3 filters that are associated to the ascending (67.5° , 45° and 22.5°) or to the descending (-22.5° , -45° and -67.5°) feature. The proposed complex second level features are important and strong characteristics of the time evolution of the spatio-temporal representation of speech (here $R(t, c)$) and are robust to noise. Another important cue and feature has also to be taken into account: Onset.

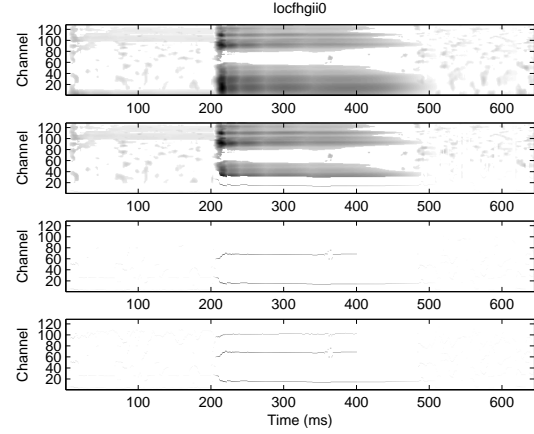


Fig. 4. Time/frequency representation $R(t, c)$ of the French vowel /i/ uttered by a female speaker (top), highest peak position estimation considered to represent F_1 (second image), detection of an important gap in $R(t, c)$ (third image) and estimation of the second peak (center of mass) F'_2 (bottom). The trajectory in the lowest frequencies represents an estimate of F_1 evolution in time and the trajectory in the highest frequencies represents that of F'_2 . The valley (gap) appears as the trajectory between F_1 and F'_2 .

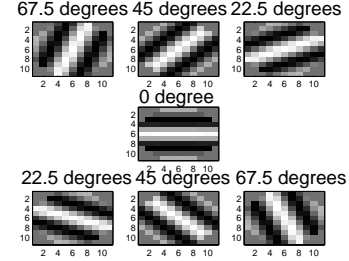


Fig. 5. The 7 orientation filters of dimensions 11x11 used to identify the trajectory tendencies. Gabor filters oriented at 67.5° , 45° and 22.5° (top row) respond to ascending trajectories. The horizontal Gabor filter (middle row) responds to stable trajectories. Gabor filters oriented at -22.5° , -45° and -67.5° (bottom row) respond to descending trajectories. Lighter zones in the figure are positive and darker zones are negative.

3) *Onset detection*: There is strong evidence that groups of neurons in the auditory system detect onsets [19], [20], [21], [22]. Also, the stimulus onset is usually robust to reverberation. As illustrated on figure 6, the output of the first level simple feature filter defined as b_{32Hz} (Eq. 1) is a good estimate of the time events at which onsets occur. The *onset* feature is the output coefficients obtained after filtering $S(t, c)$ with b_{32Hz} , and only when these coefficients are greater than the other 4 basis filtering outputs.

III. TIMING AND EVENT DETECTION

A. Timing and Event features

Features are simplified by ignoring the frequency content of the trajectories since in this work the features are not directly used in the speech recognizer. Features are used only to position frames through time. We keep the strength (magnitude) of the features and ignore the frequency values. At each instant t , 6 feature coefficients are extracted: v_{F_1} ascending, v_{F_1} descending, v_{F_1} stable, $v_{F'_2}$ ascending, $v_{F'_2}$

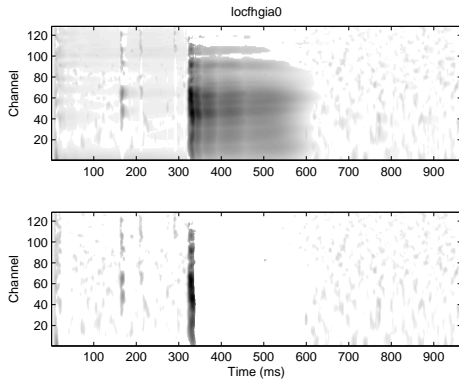


Fig. 6. Time/Frequency representation $R(t, c)$ of the French vowel /a/ uttered by a female speaker (top) and outputs of the filtering with b_{32Hz} when they are greater than the outputs from the other 4 filters (bottom). The vowel onset appears as a dark and broad frequency strip around 330 milliseconds.

descending and $v_{F_2'}$ stable. So, the nature of the speech at instant t is characterized by these 6 feature strengths. The 6 feature coefficients are inputs to six simple non-linear signal processing elements (P.E.) that are equivalent to leaky integrate and fire (LIF) neurons [23]. The signal processing element (P.E.) that responds first, labels the instant t as the strongest feature amongst the v_{F_1} ascending, v_{F_1} descending, v_{F_1} stable, $v_{F_2'}$ ascending, $v_{F_2'}$ descending and $v_{F_2'}$ stable.

Eight other processing elements are used to label the onset feature. Each P.E. integrates the onset feature signal from 16 adjacent cochlear channels (each P.E. has 16 inputs). Thus, an onset P.E. responds if the 16 adjacent cochlear channels from the onset representation to which it is connected, present an almost synchronous increase in energy.

B. The processing element model

We use a conventional model of LIF neurons [23] with a variable leaky current and a refractory period. The neuron (P.E.) emits an impulse (spike) once its membrane potential reaches a predetermined fixed threshold. The leaky current pulls the membrane potential of the neuron to the resting potential (0 here). The refractory period is modeled as a linear decrease in the leaky current after the neuron spiking. In other words, the refractory period is not absolute. The neuron is allowed to generate a spike during this period, but the leaky current is very strong just after firing and progressively decreases. In this case, it becomes significantly harder for the neuron to generate a spike since the neuron's sensitivity is decreased. With time, the leaky current also returns to its initial value.

In the case of silence or weak noise, the leaky current prevents the generation of a spike. On the other hand, when the magnitude of the feature's coefficient is sufficiently large, the trajectory that best fits the situation (ascending, descending or stable) will prompt the neuron connected to it to respond first. It is possible that more than one trajectory brings a neuron to fire. However, the first spike to be generated will come from the orientation that best fits

this particular segment, a mechanism that can be linked to Rank Order Coding [24], [2].

IV. ISOLATED DIGITS RECOGNITION

Through this section, two speech recognition systems are compared. First, a conventional fixed frame rate (FFR) Mel Frequency Cepstrum Coefficients (MFCC) and Hidden Markov Model (HMM) system is tested without modification (Reference System). Then, the processing described above is applied to create a variable frame rate (VFR) version of the same speech recognizer (designated as Proposed System).

A. Experimental Setup

Speaker independent isolated speech recognition experiments are conducted with the *TI 46-Word* database [27] that has been used in recent works like [28], [29]. The systems are developed on a desktop computer (*Intel Core2 Duo* with *Windows Vista*) under Matlab R2006a exploiting the *Voicebox* [25] and *H2M* [26] toolbox to support the MFCCs computations and HMMs respectively. The source code is available from the web site of the NECOTIS research group [39].

Each word is represented by a five-state HMM with transitions towards the two neighbors on the right. The training is performed with the *Expectation-maximization algorithm* [30]. The covariance matrix is diagonal and the observations probability density is a simple Gaussian. This simplification is often used and yields faster training [31] with reasonable performance on limited data. The recognition phase is based on the Viterbi algorithm.

A frame of twenty milliseconds length with an overlap of ten milliseconds is applied on the signal to extract twelve *MFCC* coefficients. Then the frame log energy, the Delta and Delta-Delta MFCC are combined with the MFCC coefficients and stored in the observation vector.

The training and testing data are the same as the ones provided with the *TI 46-Word* database. For all experiments, the training has been done on the clean data training set. For the reference and the proposed speech recognizers, the frame length is equal to 20ms and the same computation of the 39 dimensions MFCC based vectors is done. The only difference is that the frame of the proposed system is centered on the instants t at which at least one processing element (neuron) fires. With this strategy, the impact of choosing the position of the VFR analysis window is evaluated. Four types of noise were added: white, babble, exhibition and street noise taken from the *AURORA 2* database [32]. The Signal to Noise Ratios (SNR) was ranging from 20dB to -10dB. The impact of the reverberation was also studied for two large rooms (31x11x5 meters and 40x13x20 meters) with respective reflection coefficient values of the walls, floor and ceiling of 0.75, 0.75, 0.85, 0.25, 0.3 and 0.9. The reverberation model is based on the work of J. Allen [33] (implemented by Eric A. Lehmann). For both systems, (the reference and the proposed variable frame rate) training is always done on clean speech whether the recognition is performed on noisy or reverberated speech.

B. Experiments and Results

While testing on clean speech, the reference system has a near-perfect recognition rate (99.80%) and the proposed system has an inferior rate of 97.8%. But in noisy and reverberated conditions the variable rate frame system is more robust. With 20 dB SNR the proposed system has reasonable recognition rates (greater than 90% for all conditions) while the reference system scores decline between 57% and 66% (Table I). With 0 dB SNR, the reference system has recognition rates between 11% and 18% while our new system holds with rates between 24% and 43% (Table II).

TABLE I
REFERENCE SYSTEM: RECOGNITION RATES WITH 4 TYPES OF NOISE

Noise type	20 dB	10 dB	0 dB	-10 dB
White noise	57.75%	22.19%	10.90%	10.07%
Babble	66.72%	30.25%	14.12%	10.46%
Exhibition	60.15%	17.82%	15.77%	10.70%
Street	65.82%	21.75%	18.41%	14.72%

TABLE II
PROPOSED VFR SYSTEM: RECOGNITION RATES

Noise type	20 dB	10 dB	0 dB	-10 dB
White noise	94.34%	67.19%	38.99%	9.28%
Babble	94.41%	77.02%	25.96%	14.24%
Exhibition	90.99%	56.61%	23.68%	15.14%
Street	90.83%	79.31%	43.07%	17.11%

TABLE III
IMPACT OF THE REVERBERATION ON RECOGNITION RATES

Parameter sets	Reference System	Variable Frame Rate System
Room 1	83.36%	85.41%
Room 2	70.89%	82.14%

As shown in table III, reverberation hinders the speech recognition task and the larger the room, the harder the task becomes for the system. Again, the proposed system shows greater robustness.

V. DISCUSSION AND CONCLUSION

The frame selection algorithm focuses on important signal features. However, when the noise level increases, it becomes harder to locate the speech signal and less observation frames are generated. Compared to the reference system, the number of observation vectors used by the proposed approach drops continuously when the noise level increases and about half that number is generated at -10 dB. Given that our approach selects reliable frames, it makes sense that as noise increases, the number of reliable frames is reduced.

The processing elements (neurons) response (spike) depends on the short-term history (refractory period and adaptability of the firing threshold). They lose their sensitivity after spiking. With reverberated speech this property is very useful. Since secondary reflections of the reverberation are delayed in reference to the direct path. They have less chance

of generating a spike than the original direct path signal. Also, even if feature coefficients are high, a neuron will not necessarily spike, generating less frames than with the VFR approach that uses fixed thresholds to compare the energies and generate frames.

Our VFR system is in some sense complementary with the missing feature approach. Significant improvements in robotic speech recognition [34], [35], [36] have been obtained with the "missing feature approach" [37], [38]. In this paradigm, unreliable data from the spatio-temporal representation of speech are ignored. Our system searches and selects reliable frames while the missing feature approach discards the missing (or corrupted) information. A combination of both techniques has an excellent potential and should be investigated in future work.

In this work, features were used to position a window frame on the signal and then proceed with a conventional MFCC/HMM speech recognizer. If one is interested only in finding the reliable time events in the signal, the actual complexity of the VFR system is unnecessary. We then suggest scaling down to 24 channels. This complexity reduction is left to future work and its impact should be evaluated.

The proposed speech analysis offers a strong potential for robot audition. It has the versatility necessary for a system that needs self-adaptation. By finding and extracting the time instants at which the signal comprises specific features, voice activity and silence detection are not required. We have also shown that this approach allows the recognition of corrupted speech even when the training is done on clean speech. The recognition to the instants at which features respond strongly is more efficient than having to continuously recognize the signal regardless of its content.

To further probe our technology, we plan to design and test on larger vocabularies and with continuous speech. Because of the highly parallel architecture and of the new parallel computing technologies (like graphical processing units) a real-time implementation is conceivable as recognition is performed only at some instants and is not systematic.

ACKNOWLEDGMENT

CRSNG, Univ. de Sherbrooke, FQRNT for funding this research. S. Wood and S.K. Itaya for correcting English, S. Brodeur for constructive comments and the anonymous reviewers for excellent suggestions.

APPENDIX

Relevant properties of the brain that have inspired this work are discussed in this section.

A. Hierarchical organization of the auditory system

The auditory system comprises many nuclei with many reciprocal projections between them. Feedback and inhibition play a fundamental role in the process of sounds. Not only is inhibition active within nuclei, but also between them. This is unique to the auditory system [11]. One can hypothesize that for cochlear responses, adaptation is obtained via at least 3 loops with different time-scales (one which is internal to

the cochlea and two others that are mediated via efferents from olivo cochlear and inferior colliculus neurons). The cochlear nucleus (CN) is the first "signal" processing center where amplitude modulations (AM), ONSET, OFFSET, and possibly pitch features are extracted or enhanced. Some neurons respond stronger on transient signals while others on stable signals. In the CN, the tonotopic frequency organization of the cochlea is preserved. In the ventral lateral lemniscal, neurons seem to extract various features related to spike timings, ONSETS, and duration. The inferior colliculus (IC) is exclusively dedicated to sound processing. This has inspired our generation of the $R(t, c)$ representation. One finds the same kind of neuronal responses in higher nuclei but with a greater temporal and spatial integration. At the thalamo-cortical and cortical level, recent studies discuss the hierarchical structure of the auditory system [12]. Such organization offers the advantage of extracting a great diversity of non-linear spatio-temporal features in parallel. The neurons whose feature receptive fields most closely match the sound characteristics fire (spike) first.

B. Hierarchical organization of the visual system

It is in the first area of the visual cortex (area V1 or striate cortex) that neurons become selective to oriented targets. They respond maximally to light-dark edge flashes within their receptive field. A given visual scene appears to be encoded in the excitation of orientation selective neurons [8]. A large proportion of cortical cells are tuned to different image qualities such as direction of stimulus motion, preference to spatial frequency and temporal frequency (rate of change in contrast). In general, neurons that are selective to a particular property are located in different areas of cortex. For instance, middle temporal area (MT) contains mostly neurons that discharge selectively to the direction of the oriented moving edge no matter its color. The latter appears to be processed in area V4 regardless of the motion direction. Yet the map of the visual space is repeated in every area. It seems, however, that as the visual signal progresses to more anterior areas (for instance, IT) the receptive fields become larger and neurons may respond to various objects sharing similar attributes belonging to the same category [9]. Such hierarchical organization allows the selection and grouping of several features of complex images within one population of neurons by activation of separate sets of neurons [10]. This organization has inspired the design of the orientation filters and the 2nd type complex features.

REFERENCES

- [1] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Com.*, vol. 45, no. 4, pp. 401–423, 2005.
- [2] S. Loisel et al., "Exploration of Rank Order Coding with spiking neural networks for speech recognition," in *Proc. IJCNN*, August 2005.
- [3] J. Rouat et al., "Towards neurocomputational speech and sound processing," in *Nonlinear Speech Proc.*, LNCS, #4391, Springer, 2007.
- [4] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE J. of Selec. Top. in Sig. Proc.*, vol. 4, no. 5, pp. 798 – 807, 2010.
- [5] V. Mitra et al., "A noise-type and level-dependent MPO-based speech enhancement architecture with variable frame analysis for noise-robust speech recognition," *Proc. INTERSPEECH*, pp. 2751 – 2754, 2009.

- [6] D. M. Rasetshwane et al., "Identification of speech transients using variable frame rate analysis and wavelet packets," *Proc. IEEE Int. Conf. EMBS*, pp. 1727 – 1730, 2006.
- [7] J. Epps, "A new approach to variable frame rate front-end processing for robust speech recognition," in *Symp. SProc. and Its Appl.*, 2005.
- [8] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *J. Physiol.*, pp. 215–243, 1998.
- [9] N. Kriegeskorte et al., "Matching categorical object representations in inferior temporal cortex," *Neuron*, vol. 60, pp. 1126–1141, 2008.
- [10] S. Molotchnikoff and J. Rouat, "Brain at work: Time, sparseness and superposition principles," *Frontiers in Bioscience*, in Press 2011.
- [11] J. A. Winer and C. E. Schreiner, "The central auditory system: Functional analysis," in *The Inferior Colliculus*, Springer, 2005.
- [12] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Rev. Neuro.*, vol. 8, no. 5, pp. 393–402, 2007.
- [13] R. D. Patterson, "Auditory filter shapes derived with noise stimuli," *JASA*, vol. 59, no. 3, pp. 640–654, 1976.
- [14] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *JASA*, vol. 23, no. 2, pp. 228–248, 1961.
- [15] L. Chistovich et al., "Centers of gravity and spectral peaks as the determinants vowel quality," in *Front. of Speech Comm. Res. Ac. Press*, pp. 145–157, 1978.
- [16] A. Bladon and G. Fant, "A two-formant model and the cardinal vowels," *STL-QPSR*, vol. 19, no. 1, pp. 1–8, 1978.
- [17] R. Carlson et al., "Some studies concerning perception of isolated vowels," *STL-QPSR*, vol. 11, no. 2–3, pp. 19–35, 1970.
- [18] J. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision Research*, vol. 20, no. 10, pp. 847–856, 1980.
- [19] B. C. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 1997.
- [20] A. N. Popper and R. R. Fay, *The Mammalian Auditory Pathway: Neurophysiology*. Springer-Verlag, 1992.
- [21] J. O. Pickles, *An Introduction to the Physiology of Hearing*. Academic Press, 1988.
- [22] J. F. Brugge et al., "Sensitivity of single neurons in auditory cortex of cat to binaural tonal stimulation: Effects of varying interaural time and intensity," *J. of Neurophysiology*, vol. 32, pp. 1005–1024, 1969.
- [23] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Net.*, no. 9, pp. 1659–1671, 1997.
- [24] S. Thorpe et al., "Spike-based strategies for rapid processing," *Neural Networks*, vol. 14, no. 6–7, pp. 715–725, 2001.
- [25] Mike Brookes. (2009). [Online]. Available: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>
- [26] Olivier Capp. (2001). [Online]. Available: <http://perso.telecom-paristech.fr/cappe/h2m/>
- [27] G. R. Doddington and T. B. Schalk, "Speech recognition: Turning theory into practice," *IEEE Spectrum*, vol. 18, no. 9, September 1981.
- [28] A. Ghani et al., "Neuro-inspired speech recognition with recurrent spiking neurons," in *ICANN*, 2008, pp. 513–522.
- [29] D. Verstraeten et al., "Isolated word recognition using a liquid state machine," in *Proc. of Euro. Symp. on ANN*, April 2005, pp. 435–440.
- [30] A. P. Dempster et al., "Maximum likelihood from incomplete data via the em algorithm," *J. Royal Statist. Soc. Ser. B*, no. 1, pp. 1–38, 1977.
- [31] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2008.
- [32] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR-2000*, 2000, pp. 181–188.
- [33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, April 1979.
- [34] S. Yamamoto et al., "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory," in *IEEE-ICRA*, 2004.
- [35] J.-M. Valin et al., "Robust recognition of simultaneous speech by a mobile robot," *IEEE Tr. Robotics and Automation*, pp. 742–752, 2007.
- [36] T. Takahashi et al., "An improvement in automatic speech recognition using soft missing feature masks for robot audition," in *IEEE-IROS*, pp. 964 –969, 2010.
- [37] J. Barker et al., "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *EUROSPEECH*, 2001.
- [38] M. Cooke et al., "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, pp. 267–285, 2001.
- [39] S. Loisel, "Matlab Source code, 2011 <http://www.gel.usherbrooke.ca/necotis>,"