ROBUST SPEECH REPRESENTATION OF VOICED SOUNDS BASED ON SYNCHRONY DETERMINATION WITH PLLS

*Patricia Pelle, Claudio Estienne**

Institute of Biomedical Engineering
School of Engineering
University of Buenos Aires, Argentina

Horacio Franco

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA

ABSTRACT

We propose to include synchrony effects, known to exist in the auditory system, to represent voiced parts of the speech signal in a robust way. The system decomposes the input signal by means of a bandpass filter bank, and utilizes a bank of phase locked loops (PLLs) to obtain information on the frequencies present at a specific time. This information about the frequency distribution is transformed into a spectral-like representation based on synchrony effects. Noisy speech recognition experiments are performed using this synchrony-based spectrum, which is transformed into a small set of coefficients by using a transformation similar to that utilized for mel cepstrum features. We show that recognition performance compared to mel cepstrum features is advantageous, when measured over a range of SNR conditions, especially in the high noise level case.

Index Terms: speech features, robustness, PLL, noise, auditory system.

1. INTRODUCTION

In speech feature development, robustness to noise is an issue of great concern. It is well known that the most widely used front ends degrade notably in the presence of noise. One course of action for correcting this problem consists of developing more robust features, and one avenue to obtain such performance is to try to find features resembling how the peripheral auditory system behaves. This perceptually motivated approach has given origin to major advances in speech representation–namely, mel cepstrum [1] and perceptual lineal prediction (PLP) features [2]. Those front ends model the inner ear processing by making a frequency decomposition of the input signal into biologically inspired nonuniform bandwidth channels of constant Q, instead of a constant-bandwidth discrete Fourier transform (DFT) decomposition.

This conceptual approach enriches the representation of speech, but there are other known biological facts in the mammalian inner ear that have not yet been applied successfully to this area. One of these facts is that time-varying spectral features may be well represented by phase locked responses of the auditory nerve fibers, which are also very robust against noises added to the input signal. This observation has given origin to several attempts to represent the speech sounds. Models like Seneff's [3], Ensemble Interval Histogram (EIH) [4], or its simplified version of zero crossings with peak amplitudes (ZCPA) [5] have focused on timing information in the inner cells, and particularly on the synchronous manner in which spikes are produced, resembling almost an in-phase version of the input in the band of frequencies on which each cell operates. These models have supported, in general terms, the concept that fine time information is useful in noisy environments. In this work we present another approach that incorporates the use of synchrony to represent voiced parts of the speech signal. The system is composed of a stage that decomposes the input and extracts synchrony information in a biologically motivated way. A later stage converts the output of this first stage into a set of features suited to be applied to a conventional speech recognition system. The design of the overall system is evaluated by using the final features in a standard recognition task.

The system proposed to extract synchrony information has an architecture that we have also used in previous work, where we explored the use of synchrony-related features to represent the speech fundamental frequency (pitch) in a robust way [6, 7, 8]. A filter bank divides the signal following the common approach used in most auditory-inspired front ends, i.e., using asymmetric, overlapping and constant Q filters [9]. The filter bank outputs are then processed to obtain synchrony information by feeding them into a set of phase locked loops (PLLs). PLLs are nonlinear devices frequently used in any signal processing task related to synchrony determination, like FM demodulation, frequency multiplexing, and so on [10]. Especially interesting is the robustness of these devices in the presence of noise. The choice of this kind of device may be supported by biological evidence showing that active phenomena would be responsible for the synchronizing behavior of our auditory system [11],[12]. The choice of PLLs to obtain synchrony information is very appealing to those who are familiar with them. For example, PLLs were used by Wang and Kumaresan [13] to represent speech sounds.

There is strong agreement among the speech community about the general design and characteristics of the initial spectral decomposition of the input signal. Some agreement also exists about the importance of synchrony and time representations in the auditory system. But there is less agreement with the manner in which this information should be transformed to represent sound signals in a sense that may be used in a standard speech processing system. The approach that we follow in this work to produce such transformation is inspired by an observation pointed out in [14] and [15]. In those papers the authors mention that "synchronization" of pattern discharges of the auditory nerve "to the first formant component is particularly strong and widespread, reflecting the fact that this is the largest component in the stimulus". Similar observations are described for the rest of the formants. According to this observation, the final stage of our system is intended to display the spectral portions that are most widely signaled in the synchronization pattern. Frequencies that are often repeated at the PLL outputs, along with their variation in time, are emphasized in the spectral representation of the stimulus. The approach is similar to our earlier work, [16], but

*This work is funded by the University of Buenos Aires. Program grant: UBACYT I003.

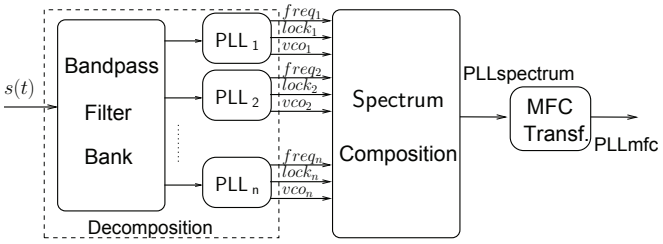

Fig. 1. System block description.

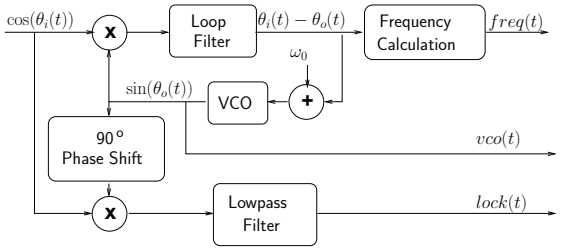

Fig. 2. Basic PLL operation.

in this case a larger number of filter and PLLs is used to obtain the spectral representation, along with other complementary validations for frequency detection that increase the robustness of the system.

The rest of the paper is divided into a section that describes in detail the filter bank and the synchronization stage based on PLLs, another section explaining the spectrum-like composition of the synchrony information, and two final sections describing the experimental setup, and results obtained as well as pending issues.

2. SYSTEM DESCRIPTION

The general system is illustrated in Fig.1. The first stage is a bank of bandpass filters that unfolds the input signal into simpler signals, each filter followed by one PLL. The outputs of each PLL are the inputs to the spectral composition stage. Finally, the spectral representation obtained is transformed using conventional transformations like the triangular filter bank and the cosine transform to convert the pseudo-spectrum into a set of PLL cepstral coefficients. In the decomposition stage, following the biological motivation, filters restrict the frequency band into which PLLs should be able to synchronize. We have chosen the kind of filters suggested by Wang and Shamma [17], which are based on biological considerations about the cochlea functioning. The general form of the filters is set according to the principle of approximately constant Q factor, covering the range between 100 Hz and 5000 Hz, linearly distributed in a mel scale. The degree of asymmetry and Q was experimentally set, using as guidance the biological descriptions of [15]. Wide filters, with a great overlap between them, allow the PLLs to be in lock with main formants, but, if the filter is too wide, it is possible to lose important inter-formant details. The degree of asymmetry is related to the number of PLLs that are phase locked to a stimulus. In our experiments we used 243 filters, $Q = 0.3$, an asymmetry factor of 0.1, and all filters are finite impulsive response (FIR) of order 2048. The signal is preemphasized before applying it to the filter bank, in order to enhance high frequencies that otherwise might be lost. An order 1 FIR filter is used, with the same coefficients as in the case of mel cepstrum calculation.

2.1. Phase locked loop operation

We used PLLs as the instrument to detect synchrony. Here, we explain PLL operation and its output signals in order to justify its use in determining synchrony. A PLL consists of a loop containing three basic blocks [10] (Fig.2): a voltage-controlled-oscillator (VCO) whose frequency is controlled by an external voltage, a phase detector that is usually a multiplier, and a low-pass filter (loop filter). The phase detector compares the phase of a periodic input signal against the phase of the VCO output, resulting in an error signal that is a function of the difference between instantaneous phases of the input ($\theta_i(t)$) and VCO ($\theta_o(t)$). This error signal is then filtered and amplified by the loop filter, and applied as a control voltage to the VCO. The VCO output is fed then to the phase detector, indicating a change in its instantaneous frequency if needed. The control voltage forces the output frequency of the VCO to vary in a direction that reduces the phase difference between VCO output and the input signal. If both phases are sufficiently close, negative feedback makes the VCO lock to or synchronize with the incoming signal. Once in lock, both VCO output and input phases are identical and, as a consequence, their frequencies are also equal. The VCO operates at an initial free-running frequency (ω_0) different from 0, which is established at the expected mean input frequency range in order to reduce time needed to be in lock. The control force applied to the VCO may be used to calculate the instantaneous frequency of the VCO, $freq(t)$. But this frequency can be considered equal to the input frequency only if the difference between the input and VCO phases is low. So an indication of the degree of lock of the PLL is provided, $lock(t)$, in order to validate the frequency indication. This signal is generated with a quadrature phase detector followed by a smoothing filter. When the difference detected by the multiplier in the main loop tends to be zero (locked condition), the output of the second phase detector tends to be maximum, and a measure of the lock-in degree of the main loop is obtained. The smoothing filter is necessary to avoid flickering of the lock indicator signal. We have used a discrete version of an analog PLL, as described in [18]. The parameters setting was made by tuning to obtain a good behavior of lock, in both clean and noisy conditions. We used a second-order loop filter, whose parameters are set to a constant $\xi = 0.5$ for all the PLLs, and a linearly varying $\omega_n/(2\pi)$ from 1 to 70 Hz. The free-running frequency of each PLL was set to the peak frequency of the corresponding filter, achieving fast in-lock state for most conditions.

2.2. Spectrum interpolation

The spectral representation is based on the PLL bank frequency outputs. The goal is to construct for each frame a vector containing some kind of statistical measure about which frequencies are mainly present in the PLL bank output, and their average drift in time within the frame. The set of these vectors can be arranged in the form of a pseudo-spectrum, similar to the power spectrum of the signal.

Some considerations should be taken into account before the PLL frequency indication can be considered a useful estimation of the true signal frequency composition. One fact is that the PLL always gives a frequency indication, even though there is not a sinusoidal signal in its input. To consider the PLL frequency as a valid measure of the input frequency, the lockin signal can be used as a primary validation measurement, as was previously mentioned. If the lockin indication is high, this means that there is a sinusoidal signal present at the PLL input, and also it indicates that both the

PLL input signal and the VCO's phases are sufficiently close that the frequency indication can be considered a true estimation of the PLL input signal frequency. But, due to the presence of the bandpass filter connected to the PLL input, a high value for the lockin signal does not allways indicate the presence of a periodic signal at the *filter* input. For example, a wide-band noise filtered by a cochlear (bandpass) filter produces a narrow-band noisy signal. This signal also behaves as a quasi-sinusoidal signal of erratic frequency centered around the peak of the filter frequency response and slowly varing amplitude. This kind of signal likewise depicts a high lockin indication, but not as response to a periodic stimulus. To overcome this problem we propose not only to verify that the lockin signal is high, but also to compare the output of each PLL with those of its neighbors. To implement this verification, the VCO's outputs of the nearest neighbors' PLLs are utilized. If the stimulus is a periodic signal, due to the great number of filter-PLLs and the superposition of frequency responses of the filters, it is likely that many neighbor PLLs synchronize to the same sinusoidal component in the periodic input, and, as a consequence, their VCO phases will also be equal. On the contrary, if the input signal is noisy, the bandpass filter outputs correspond to different signals, and also the VCO output observed in neighbors' PLLs will be different. In more detail, the first step to composing the spectral distribution of the signal consists of validation of the output frequency of each PLL$_i$, for each sample time n within an analysis frame. We consider that a PLL$_i$ is indicating a valid frequency if $lock_i(n)$ and $VCO_i(n)$ outputs meet these two conditions:

$$\begin{aligned} lock_i(n)/lock_{max}(n) &> thr_1 \\ VCO_i(n) - 1/2(VCO_{i-1}(n) + VCO_{i+1}(n)) &< thr_2 \end{aligned}$$

where $lock_{max}(n)$ is the maximum value of the lockin indicator for all PLL$_i$ at time n, and thr_1 and thr_2 are parameters of the system that must be determined experimentaly. In our case, we use $thr_1 = 0.15$ and $thr_2 = 0.1$.

When an analysis frame is completed, for each PLL$_i$ there will be a set of ordered pairs of validated frequencies and times $\{(fv_i, nv_i)\}$, whose element number may vary between a maximum equal to the total number of samples in the frame and zero. The second step is the calculation of linear regression by least squares fitting of the validated frequencies as a function of their times, for each PLL with a number of validated frequencies greater than half the number of samples in the frame. In this way we transform the set of validated frequencies into two coefficients: the validated frequencies mean $\overline{f}_i$, and the linear coefficient b_i related to the average drift of PLL frequencies with time within the frame. So, the second step consists of calculating the coefficients $\overline{f}_i$ and b_i that minimize

$$\sum_{n \in [1, nsamples]} (fv_i - \overline{f}_i - nv_i\, b_i)^2$$

for each PLL$_i$ with the required number of validated frequencies in the analysis frame. The frame width is set to 20 ms, and the frame rate is 100 frames per second.

The final step consists of composing two separate spectral descriptions with the coefficients $\overline{f}_i$ and b_i. To compose the spectral description based on the $\overline{f}_i$, a histogram of the number of occurrences of each mean frequency within a frame is calculated. This histogram is normalized by the total number of PLLs with validated frequencies in the frame. In this way, the normalized histogram is an estimate of the probability of each frequency in this frame. Formant frequencies will display a high amplitude when a high number of PLLs indicates those frequencies. We used 1000 bins to cover the range of 0 to 5000 Hz of possible values for mean frequencies. With the set of linear coefficients $\{b_i\}$ another spectral distribution is calculated, by placing at each frequency bin the average of the subset of linear coefficients b_i whose corresponding $\overline{f}_i$ coincide with this bin position. Both the histograms of mean frequencies and the linear coefficients spectral distribution are filtered with a Hamming window width of 36 bins, in order to smooth the representation. The collection of histograms of mean frequencies for each frame, and the spectral distribution of linear coefficients, may be considered as a kind of spectral representation. Finally, the two spectral representations are transformed separately, as is done with the power spectrum in the mel cepstrum case, using a 21-triangular-filter bank for frequency warping, and the cosine transform of the log of the warped representations, shifted for proper conditioning, resulting in 13 coefficients for each representation. The two sets of 13 coefficients are concatenated in a vector that we refer to as the *augmented PLLmfcc* vector. In Fig. 3 we show the power spectrum, the histogram of mean frequencies, and the linear coefficient spectral distribution for a synthetic speech signal corresponding to the emission /da/, as used in [15]. The first row shows the representation for the clean signal. It is clear that formant frequencies are the strongest in the histogram of means frequencies, while in the linear coefficients spectral distribution it can be noted that consonant /d/ has a stronger variation in time than in the case of vowel /a/. When comparing these spectra with the lower row of graphs where a white noise of 15 dB of SNR is added, it is possible to observe the robustness of the representations in the presence of noise.

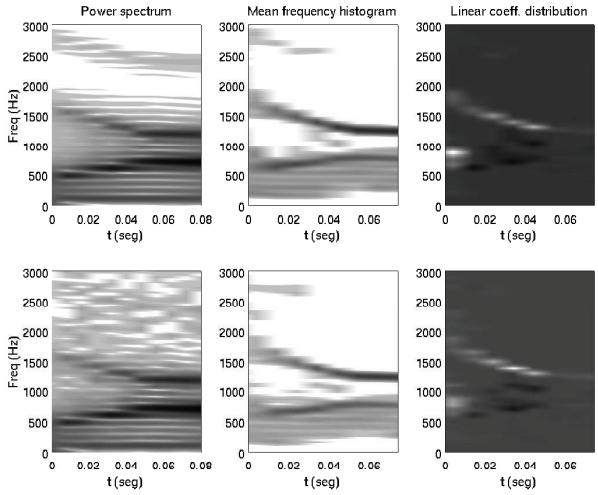

Fig. 3. Power spectrum, histogram of means frequencies and linear coefficients spectral distribution for clean signal (upper row) and SNR of 15 dB (lower row).

3. EXPERIMENTS

We used an experimental setup to test the performance of the proposed representation as described here. The proposed augmented PLLmfcc coefficients are applied to a task of vowel recognition in noisy conditions. Acoustic models for 15 vowels extracted from the TIMIT reduced phone set of 39 phonemes were trained in clean conditions. These models were used to recognize the test portion of the database, from which all other phonemes were removed. These

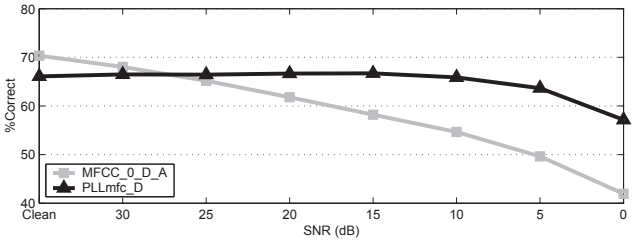

Fig. 4. Percentage of correctness in a vowel recognition task at several levels of added white noise.

tests were performed in both clean and noisy conditions where white noise was added at several levels of SNR. Noise samples were extracted from the NOISEX database, available at the Rice University Digital Signal Processing (DSP) group home page. Experimental results, in terms of correct recognition rate in percent, are compared against a standard mel cepstrum front end, using 13 coefficients, as available in the Hidden Markov Model Toolkit (HTK). Vowel acoustic models of three states, with 50 Gaussian mixtures each, are trained for mel cepstrum with Δ and $\Delta\Delta$ coefficients. For the augmented PLLmfcc coefficients only Δ coefficients are calculated, because linear regression coefficients have an implicit velocity behavior, so Δ coefficients from the augmented PLLmfcc include acceleration information. Best performance is obtained with 15 Gaussians for state. The training and testing process was also implemented in HTK. The results obtained are shown in Fig. 4.

4. RESULTS AND DISCUSSION

The obtained results show a gain in recognition accuracy for every signal-to-noise-ratio, except for those at very high or clean speech. Also it is possible to note the remarkable flat performance, unaffected by noise level, over a large range that extends up to at least 10 dB SNR. These facts allow us to conclude that the hypothesis of robustness of the proposed PLL-based representation is well founded.

Concluding, in this work we have shown useful properties of PLL-based features in the representation of speech in noisy conditions for sonorous parts of speech signals. Issues remain about the manner in which this information should be combined with other features to cover unvoiced segments.

5. REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.

[4] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 115–132, Jan 1994.

[5] D.-S. Kim, S.-Y. Lee, and R. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, Jan 1999.

[6] P. A. Pelle and M. Capelettto, "Pitch estimation using phase locked loops," in *8th European Conference on Speech communication and technology (EUROSPEECH 2003)*, Geneva, Switzerland, Sep 1-4 2003.

[7] P. Pelle, "A robust pitch extraction system based on phase locked loops," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, Toulouse, France, May 2006, pp. I–I.

[8] P. A. Pelle and C. F. Estienne, "A pitch extraction system based on phase locked loops and consensus decision," in *International Conference on Speech Communication and Technology (INTERSPEECH 2007)*, Antwerp, Belgium, Ago 27-31 2007, ISSN 1990-9772.

[9] S. Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340 – 348, 2001.

[10] F. M. Gardner, *Phaselock Techniques*. John Wiley and Sons, 1979.

[11] W. S. Rhode, "Cochlear partition vibration—recent views," *The Journal of the Acoustical Society of America*, vol. 67, no. 5, pp. 1696–1703, 1980.

[12] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," *Physiological Reviews*, vol. 81, no. 3, pp. 1305–1352, 2001.

[13] Y. Wang and R. Kumaresan, "Real time decomposition of speech into modulated components," *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. EL68–EL73, 2006.

[14] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1381–1403, 1979.

[15] M. I. Miller and M. B. Sachs, "Representation of stop consonants in the discharge patterns of auditory-nerve fibers," *The Journal of the Acoustical Society of America*, vol. 74, no. 2, pp. 502–517, 1983.

[16] C. Estienne and P. Pelle, "A synchrony front-end using phase-locked-loop techniques," in *6th International Conference on Spoken Language Processing (ICSLP 2000)*, vol. III, Beijing, China, Oct 16-20 2000, pp. 98–101.

[17] K. Wang and S. Shamma, "Auditory analysis of spectro-temporal information in acoustic signals," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 2, pp. 186–194, Mar/Apr 1995.

[18] W. Lindsey and C. M. Chie, "A survey of digital phase-locked loops," *Proceedings of the IEEE*, vol. 69, no. 4, pp. 410–431, April 1981.