Robust Speech Features based on synchrony spectrum determination using PLLs

Patricia Pelle[1], *Horacio Franco*[2], *Claudio Estienne*[1]

[1]Instituto de Ingeniería Biomédica, Facultad de Ingeniería,
Universidad de Buenos Aires**, Argentina
[2]SRI International, CA, USA
ppelle@fi.uba.ar,hef@speech.sri.com,cestien@fi.uba.ar

Abstract. In this work we propose to include synchrony effects, known to exist in the auditory system, to represent speech signal information in a robust way. The system decomposes the signal in a number of simpler signals, and utilizes a bank of Phase Locked Loops (PLLs) to obtain information of the frequencies present at each time. This information is interpolated in order to obtain a spectral-like representation based in synchrony effects, measured by the PLLs. Noisy speech recognition experiments are performed using this synchrony-based spectrum, which is transformed into a small set of coefficients by using a similar transformation as the one utilized for the Mel cepstrum features. We show their recognition performance compared to Mel cepstrum features obtained from the standard power spectrum. Some recognition improvements are obtained for the case of vocalic sounds for this approach, especially in the case of severe noise conditions.

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

1 Introduction

In the area of speech features 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 about this problem consist on developing more robust features, and one avenue to obtain such performance is to try to find features that resemble the way that the peripheral auditory system behaves. This perceptually motivated approach has given origin to mayor advances in speech representation, namely Mel cepstrum [1] and PLP (Perceptual Linear Prediction) features [2]. Those front ends model the inner ear processing by making a frequency decomposition of the input signal into biologically-inspired non-equal 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

** This work was supported by the University of Buenos Aires grant UBACYT 2008-2011, code I003

that occur in the mammalian inner ear that haven't been applied successfully yet into this area. One of these facts is that there is a synchronous representation of sounds in the outputs of the inner ear, which is 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 their simplified version Zero crossings with peak amplitudes (ZCPA) [5] have focused on the 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 that each cell operates on. These models have supported, in general terms, the concept that the fine time information is useful in noisy environments. The cost paid is higher complexity, not only in terms of computation, but also in parameters settings, which makes difficult their application as speech recognition front ends. Another important challenge in this class of biologically- or perceptually-based approaches is to decide how to process the large number of outputs typically obtained from such systems.

In this work we developed a new spectral representation incorporating the use of synchrony in the auditory representation of speech signals. This synchrony-based spectrum is then transformed in the same way that the standard DFT-derived power spectrum is transformed to obtain Mel cepstrum features, having the same dimensionality than the conventional Mel cepstrum features, but being more robust against noise interference.

The proposed new synchrony-based spectrum is obtained with a system that has an architecture that we have also used in previous works, where we explored the use of synchrony-related features to represent the speech fundamental frequency (pitch) in a robust way [6–8]. A filter bank divides the signal following a common approach used in most auditory-inspired front ends, i.e. using asymmetric, overlapping and constant Q filters [9]. The filter outputs are then processed to obtain synchrony information by feeding them into a set of Phase Locked Loops (PLLs). The PLL is a nonlinear device, often used in other areas like FM demodulation, frequency multiplexing, frequency synthesizers, etc., [10] which is known to be very robust in the presence of noise. The choice of this kind of devices may be supported by biological evidence that shows that active phenomena would be responsible for the synchronizing behavior of our auditory system [11],[12]. The choice of an active device like a PLL also allows us to make the task of parameter setting easier. The PLL obtains an internal in-phase duplicate of the main signal component present in its input. The parameters can be adjusted in order to obtain a good lock between the input and the internal representation of each PLL, and when this state is reached, further adjustments only provides marginal gain in the representation. Then, by using the frequency of the internal representation of each PLL, $freq_i(t)$, and the indication of the degree of lock obtained between the internal oscillator and the input signal, $lock_i(t)$, we compose the synchronous spectrum as the surface that interpolates the three-dimensional curves formed by $(t, freq_i(t), lock_i(t))$. As we obtain the spectrum from the synchrony output of the PLLs, we have called it a synchrony spectrum,

or abbreviated, PLLspectrum. The choice of PLLs to obtain synchrony information is very appealing to whom is familiar with them. PLLs were also used by Kumaresan [13] to represent speech sounds.

The processing approach followed in this work is also related to what we have presented in an earlier work, [14], but in this case a larger number of filter-PLLs is used to obtain the spectrum surface. In this way, in spite of the added complexity, we gain the possibility to apply another important biological observation known to occur in the representation of speech signals in the auditory system. As was pointed out in [15], an important fact that may help to determine the main features of speech in the output of the inner ear is the strong coincidence between the outputs of many cells. These persistent coincidences are present mainly at the formant frequencies, indicating that the system is devoting more resources to the representation of these important features in the spectrum. These coincidences are taken into account in this new system by using a large number of filter-PLLs, allowing a lot of redundancy in the frequencies that they are locked to. These coincidences are stronger when the energy of the signal in the pass band filter is higher, resembling in this way the biological observation. This idea of using a kind of "consensus" between several slightly different outputs has been used successfully in a previous work on pitch determination, showing a good improvement with respect to the system without this consensus reinforcement (see [8]).

In the rest of the paper we present a section with a detailed system description, another section describing the experimental set up, and finally a discussion of the obtained results.

2 System description

The system used in this work is mainly divided into three stages, as indicated in Fig.1. First, a decomposition of the input signal into simpler signals is performed by applying it to a passband filter bank, each filter followed by one PLL. This scheme was chosen because of its analogy with the cochlear functioning. After this decomposition, the outputs of each PLL are composed into a kind of spectral description by interpolation. Finally, the pseudo-spectrum obtained, or PLLspectrum, is transformed using conventional transformations like the triangular filter bank and the cosine transform, to convert the pseudo-spectrum into a set of PLL-Melcepstral coefficients.

In the decomposition stage, following the biological motivation, the 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 [16], which are based on biological considerations about the cochlea functioning. The general form of the filters are set according to the principle of approximately constant Q factor, covering the range between 100Hz and 5000Hz. The degree of asymmetry and Q was experimentally set, trying to resemble the facts described in [15]. Wide filters 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. In the presented

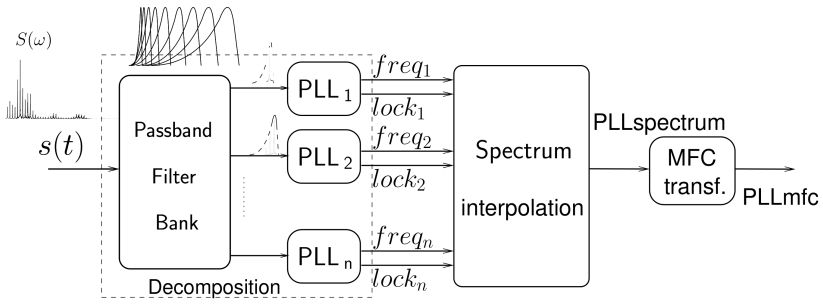

Fig. 1. System block description.

experiments we have used 243 filters, $Q = 0.1$, an asymmetry factor of 0.1, and all the filters are FIR of order 2048. The signal is preemphasized before is applied to the filter bank, to enhance high frequencies that otherwise may be lost. An order 1 FIR filter is used (one zero), as in the Mel cepstrum coefficients calculation.

2.1 Phase locked loop operation

In this work we have used PLLs as the instrument to detect synchrony. This section explains the PLL operation and its output signals in order to justify their use to determine 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 which 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 which 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 as one of the inputs to the phase detector. The VCO operates at a set frequency known as free-running frequency (ω_0). 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 to lock 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 control force applied to the VCO may be used to calculate the instantaneous frequency of the VCO, $freq(t)$. But, this frequency can be only considered equal to the input frequency, if the degree of lock between the input and VCO phases is high. 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 main phase detector output tends to zero (locked condition), the output of the second phase detector tends to be

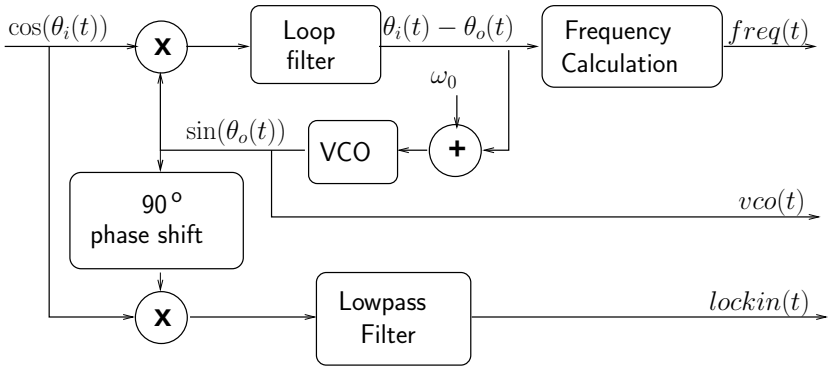

Fig. 2. Basic PLL operation.

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.

The PLLs that we have used are a discrete version of an analog PLL, as described in [17]. The parameters setting was made by tuning up to obtain a good behavior of lock, both in clean an noisy conditions. We have used as the loop filter parameters $\xi = 0.5$, $\omega_n = 0.1$, and the free running frequency was set to the peak frequency of the corresponding filter, which satisfy the condition of good lock for most conditions.

2.2 Spectrum interpolation

In order to construct the synchrony based spectrum, the interpolation between information from all PLL outputs is performed, within intervals of 0.01s. The frequency obtained at those time samples t from every PLL i, $freq_i(t)$ is discretized, and utilized as the index into a vector at which the lock indicator of each PLL, $lock_i(t)$ should be located. If more than one PLL has the same discretized frequency, their corresponding $lock(t)$ signals are summed. This kind of combination expresses that not only there is a frequency present, but that there is an equivalent degree of lock higher than the individual PLLs' lock at this frequency. With a great number of filter-PLL channels this condition is very frequent, and the indexes at which the formant frequencies appear have the highest values. This approach exploits the coincidences in frequency lock between many PLLs to represent the most salient frequencies in the input signal, in a way we use a kind of "consensus" processing to determine main speech features. Spurious isolated frequency indications not corresponding to any real sinusoidal component in the input may appear also, specially in presence of noise, but ideally the degree of lock at this frequencies should be very low.

Once this vector is formed, then it is convolved with a Hamming window to obtain an interpolated version in the frequency direction. In this work we used a grid of 1024 discretized frequencies, and a Hamming window of 200Hz

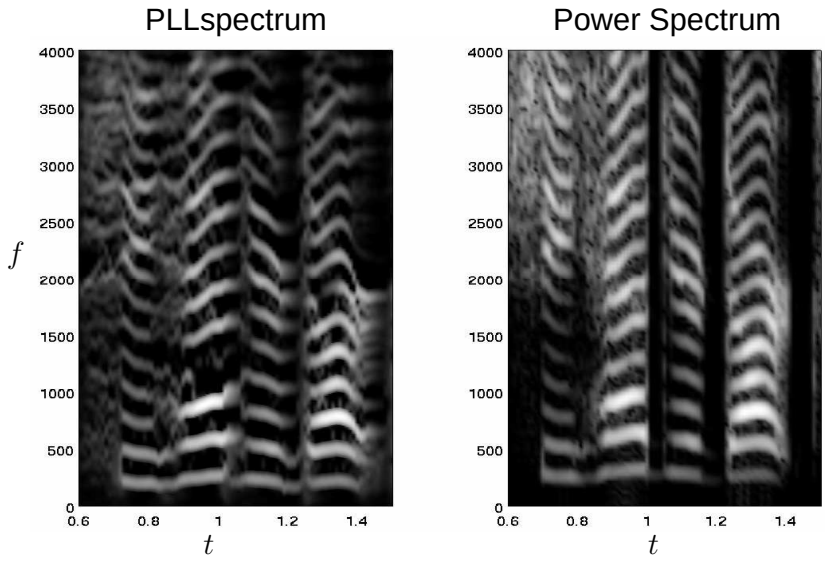

Fig. 3. PLLspectrum and Power spectrum for a clean signal.

equivalent width. In the Figure 3 we show the resulting PLLspectrum, besides the true power spectrum.

It should be noticed that both pictures look very similar, but as the PLLspectrum is a surface that represents the degree of synchrony in each PLL output used to compose the PLLspectrum, it is possible to presume that this representation should be more robust than the conventional power spectrum in presence of noises. Fig. 4 is similar to the previous figure, but with added white noise at 10 dB of SNR. It can be observed that in the power spectrum the noise is present without attenuation over the whole spectrum, particularly in the regions between harmonics. In the case of the PLLspectrum some degradation can be observed, but the regions between harmonics remain more clear.

3 Experiments and Results

The experimental set up that we have used to quantitatively test the performance of the PLLspectrum against conventional power spectrum is described in this section. As explained previously, both spectra were converted to a set of thirteen coefficients by applying the same transformations utilized in the standard Mel cepstrum case. These coefficients are applied in a task of vowel recognition in noisy conditions. Acoustic models for fourteen 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 tests were performed both in clean and noisy

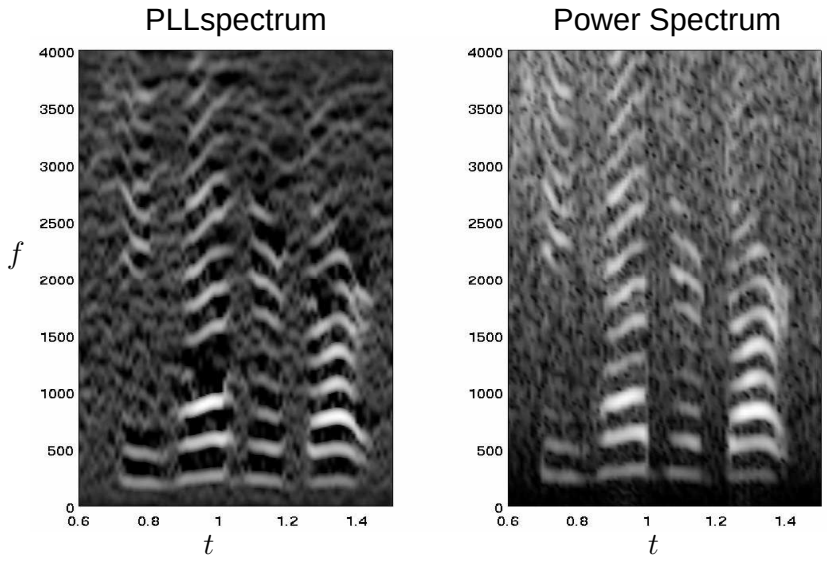

Fig. 4. PLLspectrum and Power spectrum with white noise at 10dB of SNR.

conditions where white noise was added at several levels of SNR. Noise samples were extracted from the NOISEX database examples that are available on 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 also thirteen coefficients, as available in the hidden Markov Model Toolkit (HTK). Vowel acoustic models had three states, with fifty Gaussian mixtures each. The training and testing process were also implemented in HTK. The results obtained are shown in Fig. 5 and 6, for static and static plus delta and delta-delta features, respectively.

From the results presented we may observe two facts. First the PLLspectrum has an almost constant performance for a range of moderate and medium noise levels, which allow us to conclude that the hypothesis of robustness of the synchrony-based representation is well founded. For higher noise levels, degradation is similar to the case of standard Mel cepstrum coefficients. Second, performance is better for static features than for dynamic ones, possibly implying that some details are lost in the representation, specially those that represent sudden temporal changes.

While standard Mel cepstral features degrade gradually with increasing levels of noise, one worthwhile goal would be to have a representation that is less affected by noise at different levels. We have found that this proposed spectral representation based in synchrony effects is very flexible, we can tune parameters towards this goal. For instance, the results previously presented in the above figures showed the uniform performance that can be obtained for clean and moderated noise levels. But if the usual noise levels are high, another configuration

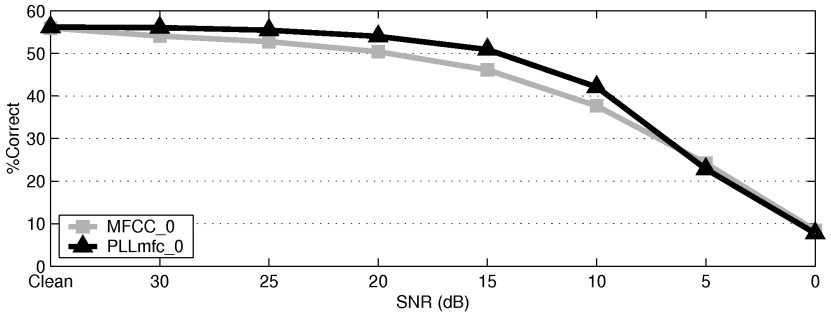

Fig. 5. Percentage of correctness for connected vowel recognition: static features at several levels of added white noise.

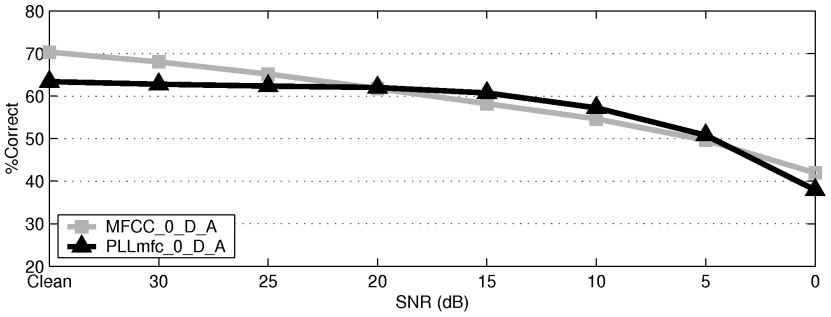

Fig. 6. Percentage of correctness for connected vowel recognition: static plus Δ and $\Delta\Delta$ features at several levels of added white noise.

of parameters that emphasizes high level of PLL lock-in produces results like those presented in Fig. 7. In this case, performance is almost constant for every noise level, including the 0dB case, which represents a great improvement with respect to the standard Mel cepstrum case, but these gains in high noise are obtained at the cost of a lower performance in clean speech.

4 Concluding remarks

In this work we have shown useful properties of biologically motivated synchrony-based features in the representation of speech in noisy conditions. Results were presented that showed that a synchrony-based spectral representation possesses robustness properties that can be used with advantage for speech recognition, specially in noisy environments. It constitutes an improvement with respect to our previous presented works, showing that implementation of this work is better. It is expected that further improvements can be applied that allows a better performance, specially in clean speech.

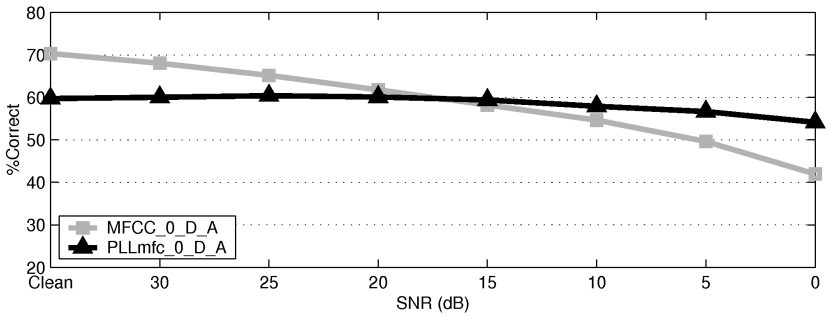

Fig. 7. Percentage of correctness for connected vowel recognition: static plus Δ and $\Delta\Delta$ features at several levels of added white noise, for a PLLspectrum with parameters optimized for high noise.

5 Acknowledgements

We thank Martin Graciarena for his helpful comments on the draft.

References

1. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 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 on Phonetics*, vol. 16, pp. 55–76, 1988.
4. O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *Speech and Audio Processing, IEEE Transactions on*, 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," *Speech and Audio Processing, IEEE Transactions on*, 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 *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, Toulouse, France, May 2006.
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. 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.
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. K. Wang and S. Shamma, "Auditory analysis of spectro-temporal information in acoustic signals," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 14, no. 2, pp. 186–194, Mar/Apr 1995.
17. 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.