Pitch-Gesture Modeling Using Subband Autocorrelation Change Detection

Malcolm Slaney[1], Elizabeth Shriberg[1], and Jui-Ting Huang[2]

[1] Microsoft Research, Mountain View, CA, USA
[2] Microsoft Online Services Division, Sunnyvale, CA, USA

malcolm@ieee.org, Elizabeth.Shriberg@microsoft.com, jthuang@microsoft.com

Abstract

Calculating speaker pitch (or f0) is typically the first computational step in modeling tone and intonation for spoken language understanding. Usually pitch is treated as a fixed, single-valued quantity. The inherent ambiguity judging the octave of pitch, as well as spurious values, leads to errors in modeling pitch gestures that propagate in a computational pipeline. We present an alternative that instead measures changes in the harmonic structure using a subband autocorrelation change detector (SACD). This approach builds upon new machine-learning ideas for how to integrate autocorrelation information across subbands. Importantly however, for modeling gestures, we preserve multiple hypotheses and integrate information from all harmonics over time. The benefits of SACD over standard pitch approaches include robustness to noise and amount of voicing. This is important for real-world data in terms of both acoustic conditions and speaking style. We discuss applications in tone and intonation modeling, and demonstrate the efficacy of the approach in a Mandarin Chinese tone-classification experiment. Results suggest that SACD could replace conventional pitch-based methods for modeling gestures in selected spoken-language processing tasks.

Index Terms: pitch, prosody, intonation, correlogram

1. Introduction

Estimating pitch can be challenging. Definitional problems include irregularities at voicing boundaries and octave ambiguities due to shifting periodicities. Engineers define pitch based on periodicity [1] and psychologists based on what we hear [9], neither of which mention the motion of the glottis as used by a speech scientist. Furthermore, computational problems are present in the face of noise and reverberation. While in many cases the pitch of a vowel is obvious, the real world is not always straightforward. The pitch of a sound is more difficult to measure as we move to address speech produced in casual, spontaneous or noisy environments.

Yet, questions about tone in languages such as Chinese, and prosodic intonation questions are often phrased as questions about pitch. We want to know the pitch of a speech signal so we can tell whether the "pitch" has gone up or down. Not only does this require us to estimate a pitch, an inherently non-linear and error-prone process, but then we compute the derivative of the "pitch." Taking the derivative of a noisy signal adds more noise.

In this paper we argue that, for some tasks, we can better answer questions about the behavior of pitch without first computing the pitch. In speech and linguistics we are often interested in what we call a "pitch gesture." We want to know whether the "pitch" goes up or down, but we don't actually care about the absolute pitch. Even with octave ambiguities and partial voicing, we see and measure clear indications of change. This change signal is more reliable, and gets us more directly to the answer we care about. We calculate pitch changes by finding many pitch candidates, as others have done, but then look at how all these candidates move over time. We never compute a *single* pitch value.

Thus in this paper we present the Subband Autocorrelation Change Detector (SACD) in Section 3, after introducing the problem and related solutions in Section 2. Section 4 describes our initial tests of the idea, and Section 5 summarizes our contribution.

2. Related Work

Pitch is inherently ambiguous. Like a Necker cube, a single sound can be perceived with more than one pitch. Shepard tones [13] are perhaps the best example. We hear a tone complex that descends in pitch forever. But how can that be? The answer is that we can often hear more than one pitch in a sound. In a Shepard tone, the pitch is continuously descending, and when one's attention is disturbed, or when the evidence for the low pitch is weak we shift our attention to a higher, more likely octave.

The root of the problem is that vocal pitch is ambiguous. One can argue that with more data and better machine learning we can find the one true pitch. But even the ground truth is problematic. Figure 1 shows the pitch-transition matrix for the Keele data [12]. These labels are computed from the laryngograph signal, and are often used to train systems and measure performance. We calculate the frame-to-frame pitch-transition matrix for the pitch labels and display the result in Figure 1. There is strong activity along diagonals one octave from the center. This suggests that one octave jumps are not rare.

The correlogram is a mature model of human sound perception [8][16]. It is based on temporal patterns, as can be measured by autocorrelation, across many cochlear filters or channels. Each channel of the filterbank corresponds to a small range of frequencies, each recorded by hair cells at one location along the basilar membrane. Within one channel auditory neurons preserve the timing of the signal, and periodicities in these firings are a robust representation of the signal. Pitch models based on the correlogram successfully pass many psychoacoustic tests [9][10][15]. The multi-channel approach is important for noise robustness.

Our work starts with the correlogram and extends it to pitch prediction using the machine-learning extensions suggested by Lee and Ellis [6]. An intermediate output of their system produces an estimate of the likelihood of 70 possible pitch classes. When combined with a Viterbi decoder, their SAcC system performs, arguably, at the limit of the accuracy of the Keele database.

Our work is close to the fundamental frequency variation spectrum idea pioneered by Laskowski et al. [4][5]. They compare the dilation of two spectrogram slices to measure pitch

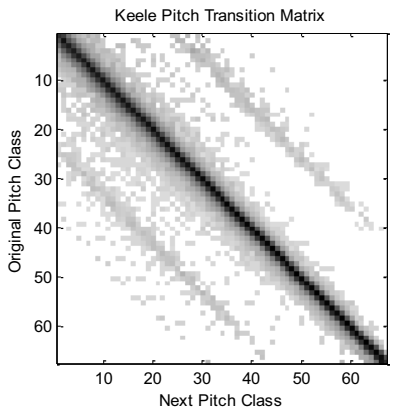

Figure 1. Pitch transition probabilities from the Keele database. Pitch is quantized into 24 classes per octave. Even this ground truth has octave jumps, as indicated by the off-diagonal lines.

changes. But a magnitude spectrogram contains the same information as a temporal autocorrelation. We extend their ideas by using multiple subbands to enhance noise robustness, use a machine-learning technique to limit the correlations to good pitch candidates, and simplify the computation by using a logarithmic-frequency scale so that we can linearly correlate frames, instead of stretching them.

In the sense that we capture many pitch candidates, our work is similar to that proposed by the RAPT algorithm, aka get_f0 [19]. RAPT uses autocorrelation of the audio waveform (no filterbank) to find a large number of possible correlation times. RAPT uses a Viterbi search to find the pitch path that smoothly goes through a collection of these pitch candidates. The Viterbi search enforces a continuity constraint that reduces the chances of an octave error. Another approach to prune the pitch candidates is called the Log-normal Tied Mixture model (LTM) [17]. The LTM approach assumes pitch is Gaussian-distributed in log space and fits three modes to a speaker's pitch distribution with means of p/2, p, and 2p. Frames whose posterior probability is higher for the first or third mode can either be "corrected" or removed. It does this without regard to the continuity of the signal, but still provides an advantage in many situations. Many more advanced models for pitch measurements are also possible [2][3][18].

In another approach, which has been used to model Mandarin tone, Lei et al. takes the RAPT pitch estimates and uses LTM to remove possible octave errors [7]. The pitch is only present when the signal is voiced. They then use interpolation to fill in the missing data, and they use two filters to give a cleaner estimate. The first filter removes the long-term trend in the pitch, as might be caused by an overall downward trend in the pitch of a sentence, or a rise at the end. They use a second filter to smooth the pitch estimates and thus give what we call a relative pitch. The combination of filters passes frequency variations between 0.66Hz and 20 Hz. They apply their ideas to tone recognition, but only as part of a larger speech-recognition system. A block diagram of their system is part of Figure 2.

The RAPT system is widely used, but as is also the case for other trackers, has difficulty at the onset of voicing, with nonmodal phonation such as creaky voice, and with noisy or reverberant signals. Post processing steps such as Viterbi searches and LTM can remove some errors. But octave errors that remain impart a lot of energy into the signal. These sharp transitions can swamp subsequent signal-processing steps.

We thus propose a more robust system, which avoids picking a single pitch. Instead we go straight to the final output, a pitch gesture.

3. System Overview

Figure 2 shows a block diagram of our system in comparison to the SAcC approach [6]. The original correlogram work gave a pitch estimate for each possible periodicity by uniformly weighting the energy across channels. Since uniform weighting is not justified, further work by Lee and Ellis [6] learn weights for different parts of the correlogram to arrive at an estimate of the pitch probabilities that best match labeled data. They implement this weighting using a multi-layer perceptron (MLP). This differentially weights the energy in the correlogram and then estimates the likelihood of each pitch class. Note, Lee first uses principal component analysis (PCA) to reduce the dimensionality of the correlogram. The goal of PCA here is to preserve the original signal, and to use a smaller number of dimensions. This makes it easier for an optimization routine to find the best perceptron weights, but doesn't affect the overall information flow.

In both SAcC and SACD there are 24 discrete pitch classes per octave, and a MLP with 70 independent outputs calculates the probability that the correlogram data includes evidence for that pitch. This results in an array of pitch probabilities for 67 frequencies from 60–400Hz on a logarithmic axis. There are three additional pitch states in this model, corresponding to unvoiced (state 1), pitch too low (state 2) and pitch too high (state 3). We trained the pitch-candidate MLP using the Keele pitch database [12].

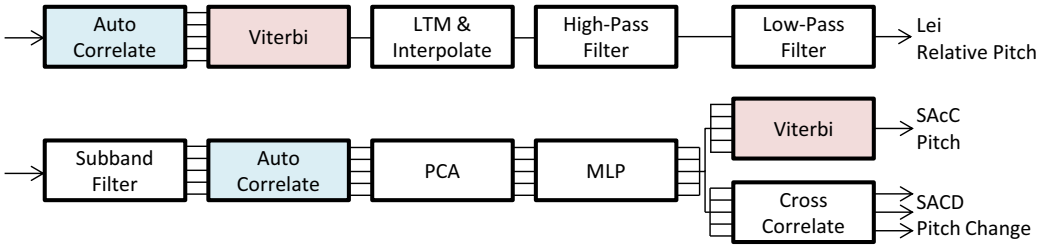

Figure 2. Three block diagrams for comparison: the baseline system by Lei et al. [7], Lee and Ellis' SAcC for estimating pitch [6], and this paper's SACD. Multiple lines are used to indicate vectors of information that are passed from stage to stage, without making an explicit decision.

The final stage of the SACD algorithm is to capture information about how the pitch-probability distribution changes over time. While there are changes from frame to frame in formants and harmonic energies, the predominant change is a vertical shift in the positions of the pitch candidates. These correspond to pitch changes. We capture these changes by correlating the pitch probabilities in one frame and the next. These changes are often small, since one unit shift corresponds to 1/24 of an octave in 10ms, but the signal is robust because it represents an average over many active pitch classes. If the probability of pitch class i is equal to p_i, then the pitch gesture (change of Δ) is computed by

$$g_\Delta = \sum_i p_i p_{i+\Delta}.$$

We implemented the subband filtering using 40 gammatone filters, as well as the correlogram calculation, using the Auditory Toolbox [14]. The MLP was implemented using Netlab [11].

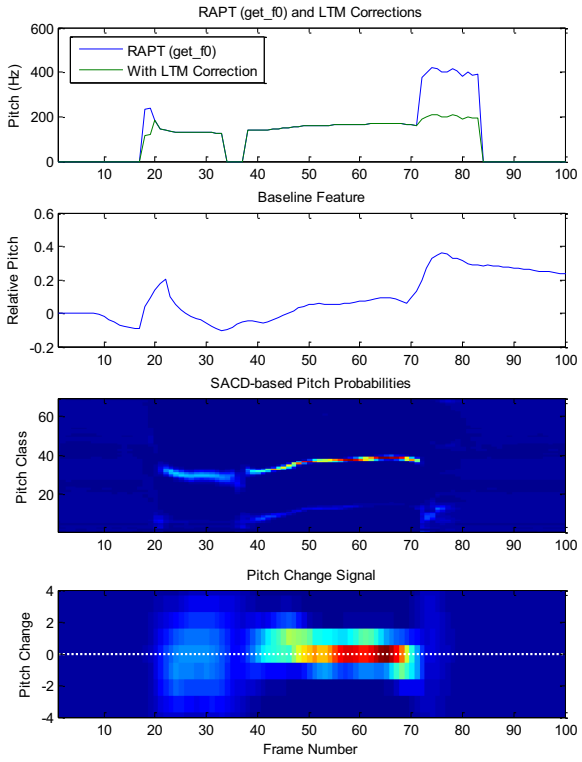

Figure 3. Baseline relative pitch vs. SACD pitch change measures of the English word "Go." The top panel shows the RAPT pitch, with and without the LTM corrections. LTM probably does the right thing around frame 80, but the right answer is not clear around frame 20. This change imparts energy into the relative pitch signal in the second panel. The third panel shows the pitch-class signal from the SACD. The fourth panel shows the pitch-change vector as a function of time. Lighter and redder colors indicate more energy at that shift. The light band above the centerline, near time 45, indicates the pitch is going up.

4. Evaluation

For a baseline, we used Talkin's RAPT code (get_f0), and re-implemented the relative-pitch feature proposed by Lei et al. [7]. This includes the LTM, the cubic interpolation across unvoiced regions, and the two moving-average filters.

Figure 3 shows a comparison of our baseline and the SACD analysis. An initial version of the SACD algorithm used an estimate of the maximum, calculated with super-resolution peak picking, as input to the classifier. But estimating the peak location can be noisy. Instead, we obtained better results by using a 5-frame moving average window to smooth the data, and then passing the 5 correlation values around 0 lag to the classifier. Thus the basic SACD pitch-change signal is a 5-dimensional vector, sampled at 100Hz.

4.1. Evaluation on Tone Classification Task

We are interested in measuring the change in pitch for a range of tasks, including for modeling intonation in natural speech and for cases in which the signal is noisy. As a first step for evaluating our approach, however, we needed a more constrained task. We chose to examine performance using a Mandarin Chinese tone recognition task, because we have large quantities of transcribed and aligned speech data. This is a simple task that involves the detection of change, rather than tasks such as emotion or speaker recognition. Lei et al. describe their system [7] with enough detail that we can replicate their algorithm and use their relative-pitch signal as a baseline. Since we could not access the tone-classification results from the prior work, we ran the published system as well as our new system on another corpus.

We started with 998 utterances from a Microsoft "Engineered Smart-Phone Database." Our data was collected by asking native Chinese speakers to read a mixed set of web-search queries and voice commands via mobile phones in their natural environment (sampling frequency of 16kHz.) The utterances were transcribed manually. The audio was then time aligned to the transcript expressed as Chinese characters. In the transcription, each Chinese character corresponds to a syllable

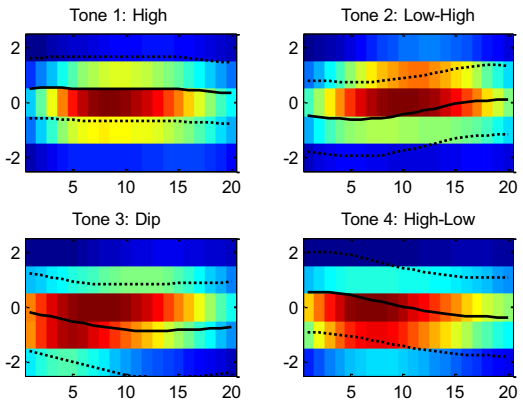

Figure 4: Average pitch gestures from the two representations for each of the four Chinese tone types. The solid line shows the relative pitch response from the baseline approach (with dashed lines indicating plus and one standard deviation from the mean.) The background images show the SACD results. All vowel examples are resampled so they are 20 frames wide before averaging.

coded with a tone; we used these tone labels in our training and testing.

From this database we extracted 5522 tone samples. The syllables with the four tone classes (high-high, low-high, dip, high-low) are on average 149ms long. Our basic evaluation metric is the four-way tone-classification accuracy, without regard to the segmental content or sentential position of the syllable. While the tone labels were generated from a dictionary, and the time alignment was machine generated, we believe this is a fair test since both systems performed the same test.

For both the baseline and the SACD system, we resampled the feature so that all test syllables had the same fixed length. This made it easier for a simple classifier to judge the information in each syllable. For display, we sorted the tones into their four classes and averaged the signal in each class. We show overall averages in Figure 4. As can be seen, both features show differences between the four classes, and the feature roughly correspond to the pitch change of each tone.

For the recognizer we used a slightly different approach. The baseline had 20 real samples, while the SACD approach has a 5-dimensional vector over time (+/-2 state changes per frame.) Thus we resampled the SACD signal so it had 4 temporal samples per syllable, so the number of variables per test for the two approaches is 4x5=20 real samples.

We trained a simple multi-layer perceptron (MLP) to classify either tone signal. For each experiment, we split the data so that a random 70% of the entire database was used for training the MLP, and the rest was used for testing. The MLP had 20 inputs, a variable number of hidden units, and 4 outputs. We judged the tone prediction as correct if the largest unit output corresponded to the correct tone.

Figure 5 shows the results as we varied the number of hidden units in the MLP from 1 to 50. In all cases the correlogram feature did better than the tone curve. This is in spite of the fact that correlogram method does not attempt to remove the long-term trend. (The lower performance when the number of hidden units is less than 10 suggests that these MLP networks don't have the necessary richness to learn the needed classifications.)

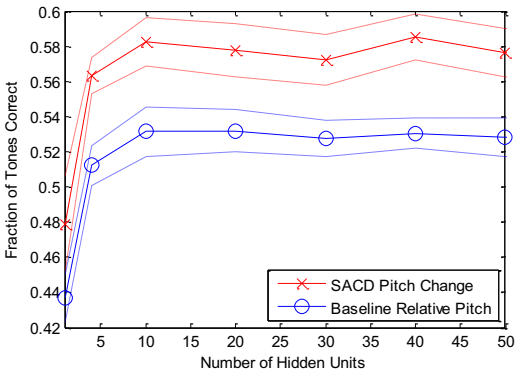

Figure 5. Four-way tone classification accuracy for the baseline and SACD features. Results are plotted as a function of the number of hidden units in the MLP. Dashed lines indicate the mean +/-1 standard error (standard deviation/sqrt(# of trials)) to give an indication of the variability in each experiment.

4.2. Evaluation with Noise

We also tested both algorithms with added white noise. As shown in Figure 6 the performance of the both approaches decline, but the gap widens as the SNR is reduced. From examination of intermediate results; we believe this is due to RAPT not producing a good pitch signal. RAPT starts to make errors as we add noise, and strong measures such as a Viterbi search and even LTM cannot compensate. The processing steps that follow when computing the relative pitch have nothing with which to work.

5. Conclusions

We have demonstrated a new system for analyzing pitch gestures. Unlike most previous approaches, we do not start with a *single* estimate of the pitch. Pitch estimates are problematic because it is difficult to find a single, best estimate, in all cases, over time, and errors are possible. When calculating the change in "pitch" these errors are compounded, so small errors become even larger derivatives. More importantly, for certain tasks we don't really care about the pitch, but rather only how it is changing.

We demonstrated the efficacy of our pitch-gesture approach in a Chinese tone-recognition task. We have presented a feature, SACD, that reflects the *change* in "pitch" over time. The feature does not start with a single pitch estimate. Instead it uses a pitch-class likelihood signal, as first pioneered in the SAcC system [6], to indicate multiple possible pitches. Even with significant amounts of noise, the SACD feature outperforms our baseline approach.

Our SACD feature is more robust to noise for two reasons. First of all, the subband analysis allows the pitch information in each channel to be analyzed separately from every other channel. A noise in one channel might obliterate the pitch in that channel, but will not affect the other channels. Secondarily, the basic pitch-class probabilities are based on a machine-learned transformation from the correlogram to the pitch class. While the MLP we use to do this transformation was trained on relatively clean (Keele) speech, additional robustness is possible with training data that matches the noise characteristics of the desired environment.

Future work should investigate the efficacy of this approach on spontaneous, reverberant and distance speech.

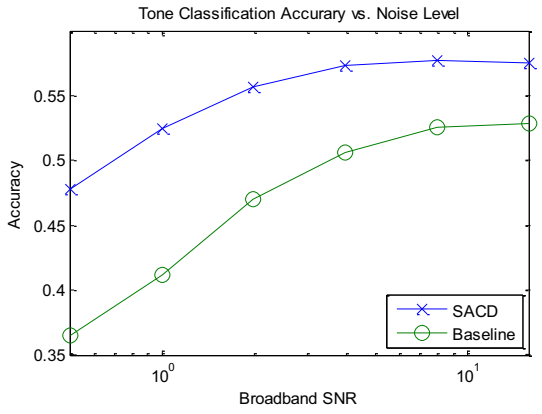

Figure 6. Four-way tone classification accuracy as a function of added noise. The SACD approach maintains its accuracy better in the face of noise.

6. References

[1] Atal, B.S., "The history of linear prediction," *IEEE Signal Processing Magazine,* vol.23, no.2, pp.154,161, March 2006.

[2] de Cheveigné, A. and Kawahara, H., "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, 111:1917, 2002.

[3] Droppo, Jasha and Acero, Alex, "Maximum a posteriori pitch tracking," *Proceedings of ICLSP'98*, pp. 943–946, 1998.

[4] Laskowski, K., Edlund, J. and Heldner, M., "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems", in *Proc. ICASSP*, pp. 5041–5044, 2008.

[5] Laskowski, Kornel; Heldner, Mattias and Edlund, Jens, "A general-purpose 32 ms prosodic vector for hidden Markov modeling" In Proc. of Interspeech 2009, Brighton, UK.

[6] Lee, B.-S. and Ellis, D., "Noise robust pitch tracking by subband autocorrelation classification," *Proc. Interspeech-12*, Portland, paper P3b.05, September 2012,.

[7] Lei, Xin; Siu, Man-Hung; Hwang, Mei-Yuh; Ostendorf, Mari and Lee, Tan, "Improved tone modeling for Mandarin broadcast news speech recognition," *INTERSPEECH*, 2006.

[8] Licklider. J. C. R. "A duplex theory of pitch perception," *Experentia* 7, 128–133. Also reprinted 1979.

[9] Meddis, R. and Hewitt, M. J., "Virtual pitch and phase-sensitivity studied using a computer model of the auditory periphery: I Pitch identification*," Journal of the Acoustical Society of America* 89, 2866-2882, 1991.

[10] Meddis, R. and Hewitt, M. J., "Virtual pitch and phase-sensitivity studied using a computer model of the auditory periphery: II phase sensitivity," *Journal of the Acoustical Society of America* 89, 2883–2894, 1991.

[11] Nabney, I. T. and Bishop, C. M., *Netlab: Algorithms for Pattern Recognition,* Springer Verlag, London, 410pp, 2002.

[12] Plante, F.; Meyer, G. F. and Ainsworth, W. A., "A pitch extraction reference database," in *EUROSPEECH*, September 1995, pp. 837–840.

[13] Shepard, Roger N., "Circularity in judgments of relative pitch," *Journal of the Acoustical Society of America* 36 (12): 2346–53, December 1964.

[14] Slaney, Malcolm, *Auditory Toolbox*, Version 2, Technical Report #1998-010, Interval Research Corporation, 1998.

[15] Slaney, M. and Lyon, R.F., "A perceptual pitch detector," *International Conference on Acoustics, Speech, and Signal Processing*, ICASSP-90, 1990, pp.357–360 vol.1, 3–6 Apr 1990.

[16] Slaney, Malcolm and Lyon, R. F., "On the importance of time—A temporal representation of sound," in *Visual Representations of Speech Signals*, eds. M. Cooke, S. Beet, and M. Crawford, J. Wiley and Sons, Sussex, England, 1993.

[17] Sönmez, M. Kemal; Heck, Larry; Weintraub, Mitchel; and Shriberg, Elizabeth, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," *EUROSPEECH*, 3, pages 1391–1394. Rhodes, Greece, September 1997.

[18] Sun, Xuejing, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.1, no., pp.I-333,I-336, 13-17 May 2002.

[19] Talkin, D., "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn & K. K. Paliwal (eds.), New York: Elsevier, 1995.