STYLIZATION OF PITCH WITH SYLLABLE-BASED LINEAR SEGMENTS

Suman Ravuri and Daniel P.W. Ellis

LabROSA, Dept. of Electrical Engineering
Columbia University, New York NY 10027 USA
ravuri@eecs.berkeley.edu, dpwe@ee.columbia.edu

ABSTRACT

Fundamental frequency contours for speech, as obtained by common pitch tracking algorithms, contain a great deal of fine detail that is unlikely to hold much perceptual significance for listeners. In our experiments, a radically reduced pitch contour consisting of a single linear segment for each syllable was found to judged as equally natural as the original pitch track by listeners, based on high-quality analysis-synthesis. We describe the algorithms both for segmenting speech into syllables based on fitting Gaussians to the energy envelope, and for approximating the pitch contour by independent linear segments for each syllable. We report our web-based test in which 40 listeners compared the stylized pitch contour resyntheses to equivalent resyntheses based on the original pitch track, and also to pitch tracks stylized by the existing Momel algorithm. Listeners preferred the original pitch contour to the linear approximation in only 60% of cases, where 50% would indicate random guessing. By contrast, the original was preferred over Momel in 74% of cases.

Index Terms— Speech analysis, Speech processing, Piecewise linear approximation

1. INTRODUCTION

Voice pitch or intonation is a major component of the non-lexical 'prosodic' content of speech, and carries important information relating to phrasing, utterance type (question vs. statement), stressing particular words, etc. However, a pitch tracker will return estimates of fundamental frequency (f_0) as a function of time that vary on a millisecond scale, whereas linguistic analysis suggests that pitch information is organized only at the level of syllables or words (hundreds of milliseconds) [7]. This paper is concerned with simplifying, or *stylizing*, the raw pitch track derived from a signal to see to what extent detail can be removed without affecting the information and/or quality of the speech. A successful stylization scheme could have applications in reducing the data rate in speech coding, as well as pointing to the kind of internal representation or processing of speech employed by listeners.

First author is now with EECS Department, University of California, Berkeley. This work was supported in part by the NSF (grant IIS-0535168).

Momel [4] is one example of an approach to this solution, and its inclusion in the widely-used PRAAT software [1, 3] perhaps makes it the de-facto standard for pitch contour stylization. The algorithm is fairly simple: it finds minima and maxima boundary points within a certain time window and fits a 2^{nd}-order spline function to the boundary markers. Despite being a much smoother representation of the original pitch track, human subjects are reported to find the stylized contours perceptually unimpaired – direct evidence that listeners are not particularly sensitive to subtle details in voice pitch contour.

Based on the idea that pitch gestures are perceived as properties of words (or at the finest scale, the stressed syllables within words), we wished to experiment with stylizations that first segment speech into syllables, then describe the pitch contour with a few parameters per syllable. In pilot experiments, we found that using just one pitch parameter per syllable – a constant, average pitch – significantly impaired the perceived naturalness of the speech, but, to our surprise, a two-parameter model – average pitch plus a constant pitch slope – afforded resyntheses that were frequently difficult to distinguish from the originals, despite being a highly stylized, and rather implausible, contour.

In order to easily manipulate a number of speech parameters, we use the STRAIGHT analysis-synthesis tool [5]. It is based on a source-filter model and achieves highly natural speech synthesis. Using STRAIGHT allowed us to separate pitch track from spectral envelope, optionally substitute stylized versions of the pitch track, then resynthesize speech of very high perceptual quality to use in listening experiments.

The next section describes our pitch stylization system, and section 3 describes our evaluation of both syllable segmentation and perceptual quality of the stylized pitches. We discuss some future directions in section 4.

2. OVERVIEW

Our pitch stylizer has two separate parts: a syllable segmenter to segment pitch by syllable, and a pitch stylizer to model the pitch within each syllable as a linear segment. Figure 1 illustrates the entire process, and is described in more detail in the subsections below.

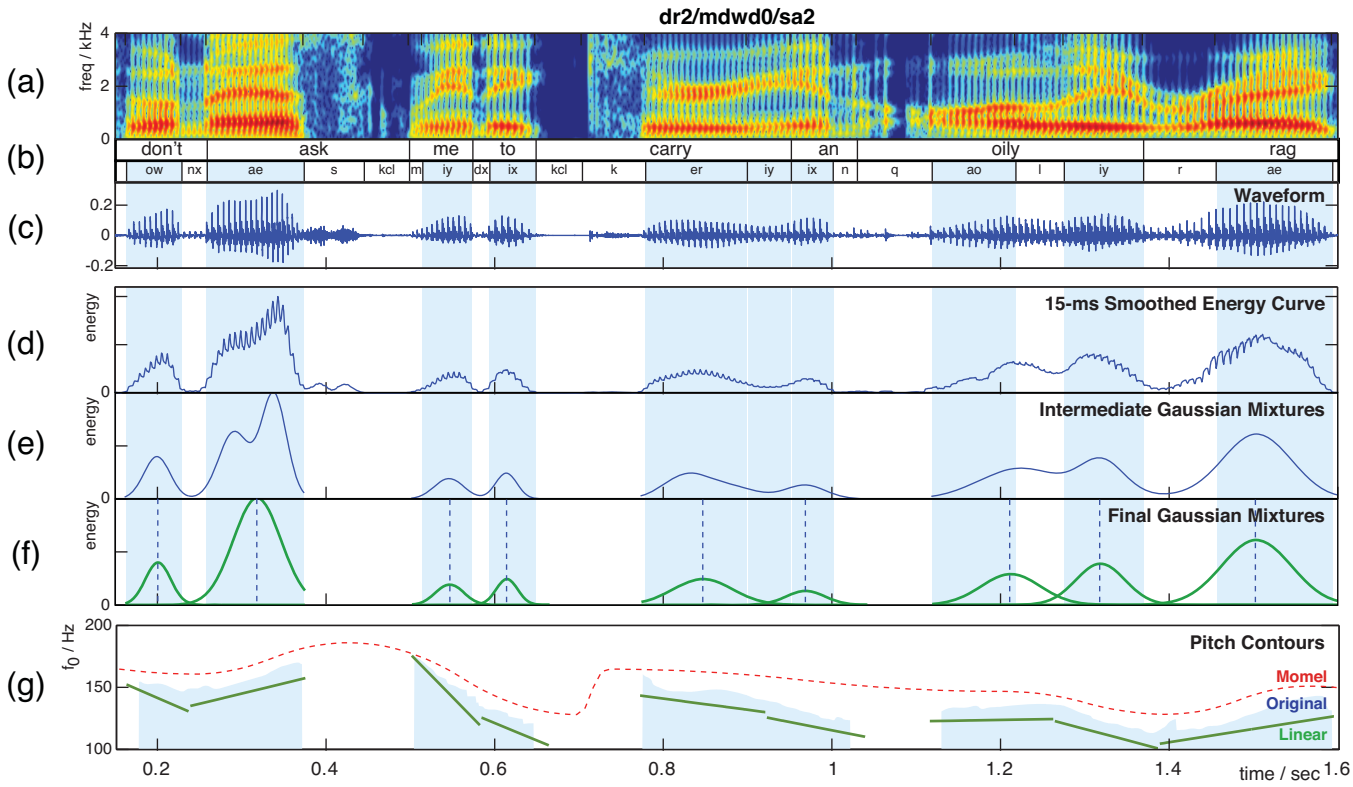

Fig. 1. Overview of the syllable segmentation and pitch stylization system. (a) Spectrogram of the original speech, drawn from the TIMIT corpus [2]. (b) Manual word and phone transcripts. Shaded backgrounds indicate vowel phones, taken as syllable centers. (c) Original waveform. (d) Energy envelope, smoothed with a 15 ms window. (e) First stage Gaussian modeling (overfit with more than one Gaussian per syllable; the sum of all Gaussians is plotted). (f) Second stage Gaussian modeling; each Gaussian is separately plotted with a dotted line at its center, indicating a syllable detected by the system. (g) Original pitch contour (shaded), plus Momel (dotted) and linear (solid) stylizations. The Momel curve has been displaced up by 10 Hz, and the linear curve down by 10 Hz, to aid clarity.

2.1. Syllable Segmentation

An integral part to the pitch modeling scheme is the syllable segmentation since the pitch curve is modeled on syllabic units. We start with the energy curve of the original waveform, smoothed with a 15 ms Hanning window. The goal is to approximate this curve with a sequence of one-dimensional Gaussian curves, with each Gaussian representing (and thus defining) the energy due to a separate syllable. Gaussian mixtures can be fit to any curve using the Expectation-Maximization (EM) algorithm, and the fit can be made arbitrarily close by increasing the number of Gaussians. The challenge is to decide how many Gaussians are needed to fit the variations in energy due to syllables, but not sub-syllabic structure. Our solution was a two-stage Gaussian modeling approach, which first uses a relatively large number of Gaussians to approximate the raw energy curve, then makes a second Gaussian mixture approximation to the result of the first stage. The steps in the procedure are as follows: (1) Segment the energy curve into 250 ms segments and determine the number of local maxima (N) in each frame. (2) Fit each segment's energy curve with a number of Gaussians moving from $N-3$ to $N+3$ until the mean-squared error improves by less than 10% for two successive increments. (3) Count the number of local maxima in the combined Gaussian energy curve of (2) that are also the peak values within a 250 ms window (N'). (4) Model the energy curve from (2) using N' Gaussians. Overfitting in steps (1) and (2) causes no problems in this segmentation strategy as the extra Gaussians are unlikely to produce extra maxima in step (3). Syllable boundaries are then taken as the point at which adjacent Gaussians cross.

Panels (d) through (f) of figure 1 illustrate this procedure. Notice that although the /ae/ in "ask" is initially modeled with two maxima in the first stage (panel (e)), these are collapsed to a single Gaussian in the second stage (panel (f)).

2.2. Pitch Model

Once the pitch track is segmented by syllable, the pitch track within each syllable is modeled as a straight line segment,

discarding the microprosodic information. In order to ensure robustness against bad pitch estimates, the linear fit is weighted by the energy curve of the original speech. Empirically we observed that while the pitch tracker we used correctly identify unvoiced sections in most circumstances, at times it would make spurious pitch estimates as voiced sections were transitioning to unvoiced sections. This could cause the edge of a pitch track to change rapidly and seriously impact the fit. These sections, however, have very low energy in comparison to voiced sections; hence weighting the linear regression based on the energy curve discounted these problems.

We applied this stylization to the outputs of a number of pitch trackers. The most common gross pitch tracking errors that we observed, pitch halving and doubling, are not corrected through stylization unless they occur in regions of very low energy. Unlike pitch stylization procedures that attempt to match the entire pitch curve of a phrase, however, because in this model each syllable is fit independent of all others, a bad pitch stylization is limited in impact to the syllables in which the bad pitch estimate occurs.

Panel (g) of figure 1 shows the original pitch tracker output (shaded) and the per-syllable linear fits (solid lines, offset down by 10 Hz for clarity). Also shown for comparison is the output of the Momel algorithm (dotted, offset upwards) which connects across unvoiced regions.

3. EVALUATION

Syllable segmentation and pitch stylization were evaluated separately, using an objective measure for the former, and subjective listening tests for the latter.

3.1. Syllable Segmentation

To test the accuracy of the syllable segmenter, we ran the algorithm on the TIMIT training database (with 50 speakers and roughly 5000 speech samples) and evaluated our syllable segments against the TIMIT phoneme labels. We also compared our algorithm to a minima-based syllable extractor (similar to [6]) which looked for minima with in 40ms windows and identified those minima as syllable boundaries. However, the TIMIT data does not directly provide ground truth syllable segmentation, only the hand-marked phone labels. Definitive syllable boundaries are notoriously elusive, as illustrated by the words "..carry an.." in figure 1, which could be argued to constitute either two or three syllables.

Our approach was to consider each vowel phoneme as a syllable nucleus, and to sidestep the problem of defining precisely where the boundaries between successive syllables occur. We considered a syllable to be correctly detected if boundaries are placed before the beginning of the corresponding vowel and after the end of the vowel nucleus. In figure 1, "..carry an.." is thus considered as three syllables in the ground truth (since it has three vocalic phonemes), but only two of these are correctly bracketed by Gaussians.

With this metric, the GMM syllable segmenter achieved an accuracy of 78% while the minima classifier correctly segmented vowel nuclei 70% of the time. The downside to using the GMM syllable segmenter, however, was that it was two orders of magnitude slower than the minima-based classifier. We should also note that the speakers in the TIMIT database pronounce their words quite clearly, leading to easier segmentation of syllables. The accuracy numbers for the GMM classifier and the minima classifier are likely to be significantly worse in conversational speech.

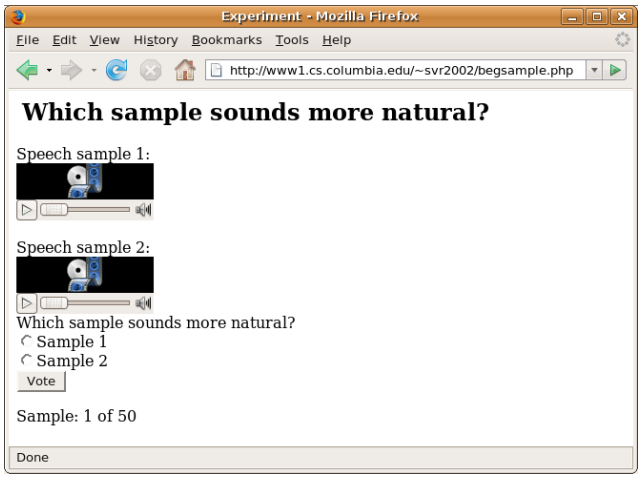

Fig. 2. User interface of the web experiment.

3.2. Human Evaluation of Pitch Model

In order to evaluate the naturalness of synthesized speech based on the pitch stylization model, we created a web experiment in which listeners are given two versions of the same utterance from the TIMIT database. Using the same STRAIGHT resynthesis for both, one is generated from the original pitch contour, while the other uses a stylized pitch. Subjects are asked to choose the more natural-sounding speech sample. Figure 2 shows a sample screen from the web experiment.

In the experiment, 20% of trials used pitch stylizations by Momel instead of the linear model. Both stylizations were based on the pitch track produced by PRAAT [1], which was also the source of the Momel implementation. For about 3% of the TIMIT samples, Momel produced pitch tracks with large oscillations giving peak f_0 values above 10 kHz; these tokens were excluded from the experiment. For speed of processing, we used ground-truth syllable boundaries inferred from the TIMIT transcription.

40 subjects participated in the evaluation, totaling over 1,500 judgments. Table 1 shows that individuals are able to distinguish the original from the linear stylization 59.6% of the time compared with 74.0% for Momel. While some in-

Table 1. Percentage of times subjects were able to correctly identify the original speech sample from the stylized one. The 'Preference' column shows the number of subjects, of the subset who heard both stylizations, who more often preferred each stylization type.

System	Ability to distinguish	Preference
Linear	59.6%	26
Momel	74.0%	8

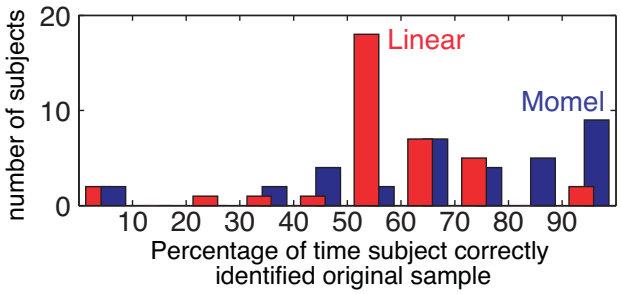

Fig. 3. Histograms of the proportion of trials in which the original was considered more natural, broken down by subject. 100% indicates perfect preference for original, and 50% indicates an inability to distinguish the versions. Light bars are the Linear model, dark bars are Momel.

dividuals are clearly able to distinguish between the original and the linear stylization, the results show that this is only for a minority of cases. Figure 3 is a histogram which shows the distribution of per-user average preference for the original over each of the stylizations. This graph shows that a majority of subjects correctly identified the original speech sample over the linear stylization only 50-60% of the time (i.e. close to guessing), while for Momel the highest concentration of subjects scored between 90-100% or 60-70%. The linear stylization is clearly perceived as more natural than Momel.

4. CONCLUSIONS AND FUTURE WORK

While the linear stylization seems to produce speech samples indistinguishable from the original a large percentage of the time, listeners are still able to discern differences between the original and stylized tracks in at least some of the samples. Looking back at figure 1, we notice that while most of the pitch track seems to be well fit by a syllabic linear segments, a few syllables such as the first half of "carry" show significant deviations from linearity. We have experimented with fitting 2^{nd}-order functions when the mean-squared error between the linear stylization and original pitch track reaches a certain threshold, but have not evaluated this approach. Speakers in TIMIT database tend to pronounce words clearly using relatively wide pitch variation; perhaps the reduced intonation of conversational speech would affect subjects' ability to distinguish between pitch stylizations. Finally, it would very interesting to perform these tests using a tonal language such as Mandarin to determine whether native speakers of such languages have a different sensitivity to pitch contour.

In conclusion, the results show that native English speakers are to a large extent insensitive to microprosody within a syllable. Even a crude fit, such as the linear regression we used, seems adequate for realizing natural speech, with listeners able to distinguish stylized from original pitch in only a small proportion of cases. We believe this has important implications for representations of speech, and the exploitation of intonation information in speech systems.

Examples of the pitch stylizations described in this paper can be heard at `http://labrosa.ee.columbia.edu/projects/pitchcontour/`.

5. REFERENCES

[1] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.3.14), 2005. Computer program, available: `http://www.praat.org/`.

[2] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993.

[3] D. Hirst. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In *Proc. Int. Conf. Phonetic Sci. XVI*, Saarbrücken, 2007.

[4] D. Hirst and R. Espesser. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de phonétique d'Aix*, 15:71–85, 1993.

[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 3-4(27):187–207, 1999.

[6] P. Mermelstein. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am.*, 58(4):880–883, 1975.

[7] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A standard for labeling English prosody. In *Proc. Int. Conf. on Spoken Language Proc.*, pages 867–870, Banff, Canada, 1992.