Synthetic F_0 Can Effectively Convey Speaker ID in Delexicalized Speech

Eric Morley, Esther Klabbers, Jan van Santen, Alexander Kain, Seyed Hamidreza Mohammadi
Center for Spoken Language Understanding, Oregon Health & Science University
Portland, OR 97239, USA

Abstract

We investigate the extent to which F_0 can convey speaker ID in the absence of spectral, segmental, and durational information. We propose two methods of F_0 synthesis based on the Linear Alignment Model (LAM) [2]: one parametric, the other corpus-based. Through a perceptual experiment, we show that F_0 alone is able to convey information about speaker ID. We find that F_0 synthesized with either LAM-based method conveys speaker ID almost as effectively as natural F_0.

Index Terms: F_0, prosody, speech synthesis, speaker identity, recombinant synthesis

1. Introduction

Pitch, or more precisely its physical substrate F_0, is a fundamental component of prosody that conveys a great deal of paralinguistic information, including speaker identity [1]. Here we examine speaker-specific F_0 synthesis with an eye towards creating synthetic voices that sound like specific target speakers.

We first present a simple algorithm that decomposes natural F_0 curves into the component curves posited by the *Linear Alignment Model* (LAM) [2]. We then present two methods for synthesizing speaker-specific F_0 contours, both of which are instantiations of the LAM, and therefore use the output of the F_0 decomposition algorithm. The first method synthesizes F_0 parametrically, following the Simplified Linear Alignment Model [7]. The second method recombines natural F_0 component curves, and builds on the work of van Santen et al. [3, 4]

Next, we investigate how effectively F_0 can convey speaker identity in the absence of spectral, segmental, and durational information by conducting a perceptual experiment, in which we compare both natural and synthetic F_0 contours. This experiment addresses two questions: (1) How well can listeners identify speaker identity using F_0 alone, and (2) Do the proposed methods of F_0 synthesis actually capture and convey speaker identity? The first of these questions serves to illustrate an upper limit to how effectively any model of F_0 can convey speaker identity.

While it appears that certain characteristics of F_0, for example the mean, are necessary, or at least help to convey speaker identity, we are not aware of any investigations testing whether F_0 is sufficient to convey speaker identity to human listeners. The results of our study are highly relevant for determining how to best evaluate synthesized F_0 contours, and more importantly, models for F_0 synthesis.

2. Background

2.1. F_0 and Speaker Identity

The performance of automated speaker-identification and -verification systems can benefit from features extracted from F_0 [1]. It seems likely that humans make use of F_0 in identifying speakers as well, but we are unaware of any studies examining the extent to which F_0 is necessary, or whether it is sufficient, for humans to identify particular speakers. Klabbers et al. have found that in synthetic speech, speaker identity is more effectively conveyed with the addition of prosodic information from the target speaker than with spectral information alone [5]. This suggests that a synthetic voice that models a target speaker will sound more like that speaker if it can effectively model its F_0 contours.

This material was supported by the National Science Foundation Grants No. IIS-0964468 and IIS-0905095.

2.2. The Linear Alignment Model

The LAM is a model of F_0 first introduced by van Santen and Möbius [2]. The LAM posits that utterances are divided into left-headed *feet*. A foot is defined here as an accented syllable (i. e. a syllable bearing pitch accent) followed by any number of unaccented syllables. It occurs immediately before the next accented foot or phrase boundary, whichever is first. All of the phonetic material between a phrase boundary and the next accented syllable is called the *anacrusis*. All of the phonetic material between phrase boundaries is called the *phrase*. Example (1) illustrates each of these divisions: accented syllables are underlined; commas indicate phrase boundaries; feet are indicated with square brackets; the anacrusis is indicated with curly brackets; and phrases are indicated by angle brackets:

(1) ⟨ {The} [money is on the] [table] ⟩, ⟨ [not the] [counter] ⟩.

Like the Fujisaki Model of F_0, the LAM is a superpositional model, which means that F_0 contours are viewed as the combination of *component curves* [6]. Specifically, the LAM asserts that F_0 contours are the summation of the following three component curves:

Phrase curve (PC) captures the overall trajectory of F_0 over a single phrase

Accent curve (AC) captures the rise and fall of F_0 associated with a single foot; starts at 0 Hz, rises monotonically to a maximum, then falls monotonically to 0 Hz

Segmental perturbation curve captures the spike in F_0 that typically accompanies certain phones, for example /p^h/; ignored in this paper

Besides perturbation curves, the key difference with the Fujisaki model is how accent curves are computed, specifically how they are aligned with feet. This alignment is given by a linear weighted combination of the durations of groups of phonemes within the foot. These weights are called *alignment parameters*. For the purposes of the present paper, we assume that the PC is piecewise linear, with a change in slope located at the beginning of the final foot in the phrase. Note that there is a single AC per foot because each foot contains exactly one accented syllable. Figure 1 shows a natural F_0 contour and its component curves.

Van Santen et al. proposed a wavelet-based method for decomposing F_0 contours into component curves, citing two reasons for not taking a simpler approach based on local minima: (1) ACs may overlap, in which case the local minima will not be on, but rather above, the PC; and (2) the discontinuity in the PC at the start of a phrase-final foot [7]. We make the crude assumption that our corpus of F_0 contours either does not contain overlapping ACs, or that if it does, the overlap is so minor

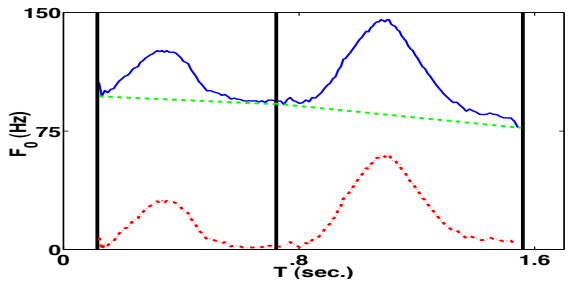

Figure 1: Decomposed F_0 contour: F_0 in blue (solid); phrase curve in green (dashed); accent curve in red (dashed and dotted); and foot boundaries in black (vertical)

that our algorithm, which is based on local minima, will still yield decompositions that are accurate. Manual inspection of the F_0 contours decomposed in our investigation suggests that the method we propose works well enough to decompose the F_0 contours in our corpus (introduced below). Regardless, if this simple decomposition algorithm performs poorly, one can easily substitute any other F_0 decomposition algorithm [8, for example] for the one described here, and still use the F_0 synthesis methods we describe below without any changes.

2.3. CSLU F_0 Protocol

Our experiments used the CSLU F_0 Protocol (FP), a set of all-sonorant utterances designed to elicit F_0 contours produced with various linguistic and prosodic features. In addition to the phonetic content of each utterance, the FP also specifies pause locations. Furthermore, it prescribes which syllables must bear pitch accent, or in terms of the LAM (Section 2.2), it indicates the location of each foot. Finally, each utterance in the FP has a target word that is spoken with a prescribed degree of emphasis.

The various features controlled for in the FP, as well as their possible values, are shown in Table 1. Example (2) is an utterance from the FP, and is annotated using the same markup as was used in Example (1), with the target word in small caps.

(2) [Will we] [really know] [MARIO], [when we're in] [Maine?]

We extract F_0 from the utterances in the FP with Praat, and correct obvious F_0 tracking errors by hand. There are two benefits to using automatically extracted F_0 contours, rather than hand corrected contours based on glottal closure instances (GCIs) in this experiment. First, manual correction of GCIs is extremely labor-intensive, so it would be preferable to develop intonational models from automatically estimated F_0 values. Second, automatically extracted F_0 contours are smoother than manually extracted ones, and therefore they are less likely to contain large jumps in F_0 during areas of mild creaking. The F_0 decomposition algorithm we present in Section 3 is highly sensitive to local minima, and so may actually perform better with automatically extracted F_0 contours than hand-corrected ones.

We have two variants of the FP: a full version containing 229 utterances, and an abridged version containing 61. We have collected the full FP from five adult speakers (two female, three male), and the abridged FP from eight adult speakers (five female, three male). All of the speakers are native speakers of English. In this experiment, we model the F_0 contours of six of these speakers, three male and three female, using only the abridged FP, and ignoring utterances with higher levels of emphasis. Table 2 shows the means and standard deviations of F_0 for these six speakers.

Feature	Values
contrast type	none; contrastive
sentence type	declarative; wh-question; yes / no question
number of syllables in target word	1; 2; 3 or more
phrasal position of target word	initial; medial; final

Table 1: Features and values specified in the FP

3. Automated F_0 Decomposition

We will now describe a fast method of decomposing F_0 contours into phrase and accent curves. We also discuss how we use the decomposed curves yielded by this method to train speaker-specific models of F_0 synthesis. We store speaker-specific parameters estimated from the decomposed curves for the parametric model (PARAM), and the component curves themselves for the second model (RECOMBINANT). Both of these speaker-specific models for F_0 synthesis will be explained in Section 4.

The decomposition algorithm requires two types of input per utterance: (1) *reliable* F_0 values, and (2) foot boundary timestamps and labels (foot/anacrusis). In this paper, F_0 values were visually inspected, and foot segmentation was manually corrected. It is critical that F_0 values used with this algorithm be as reliable as possible, as spurious local minima or maxima can easily lead to erroneous parameter or component curve estimates. As a result, if hand-corrected or visually inspected F_0 values are unavailable, we recommend discarding less-reliable points on the F_0 trajectory (for example by keeping only those points where the product of the voicing flag and energy exceeds a threshold) before using these curves with the proposed decomposition algorithm.

First, we estimate the PC as follows: given the (reliable) F_0 and the foot timestamps, we find local minima in F_0 within a small window surrounding each foot boundary[1]. We assume these local minima to be points directly on the PC (i. e. AC=0 Hz at these locations). As previously mentioned, if there are overlapping accent curves, these local minima will actually be above rather than on the PC.

We then model the PC by fitting a low-order polynomial to the observed points on the PC. The order of this polynomial depends upon the number of local minima we find, which is itself determined by the number of feet in the phrase, and whether there are reliable F_0 values around each foot boundary. If the quality of F_0 is low, there may not be any reliable F_0 values near a given foot boundary. With two minima, we approximate the PC as a linear curve, with three or four as quadratic, and with more than four as cubic. Furthermore, we weight the observed points such that the estimated PC must pass through the observed points at the beginning of the first foot and the end of the last foot in each phrase. Finally, we subtract the PC from F_0 to recover the ACs.

At this point, the F_0 curve has been decomposed into phrase and accent curves. We can store the component curves themselves, or estimate parameters from them, depending upon the synthesis method. To store the component curves for RECOMBINANT, we store the raw ACs, and the polynomial coefficients that describe the PC along with information about the foot specified in the FP (number of syllables, phrasal position and type, emphasis level and type).

For the parametric model, we estimate the following six parameters (all valued in Hz): the height of the PC at the (1) begin-

[1] If an anacrusis is present in a phrase, we ignore the left boundary of the anacrusis, because it is not a true foot boundary.

Gender		Mean (Hz)	Std. (Hz)
Female	1	133	18
	2	182	24
	3	203	31
Male	1	94	9
	2	107	19
	3	114	20

Table 2: Summary statistics of 6 speakers in the FP

ning; (2) end; and (3) inflection point at the start of the last foot in the phrase; and the maximum height of the ACs in phrase (4) -initial; (5) -medial; and (6) -final positions. We take the median value of each parameter over all of a target speaker's utterances to estimate them for that speaker. We note that no alignment parameters are stored; AC alignment is computed during synthesis using further simplifications of the LAM.

4. F_0 Synthesis

4.1. Parametric Synthesis

To synthesize an F_0 trajectory parametrically, we first create the phrase curves using the three parameters that describe its height at the start point, end point, and inflection point at the start of the last foot in the phrase. If there is more than one phrase, we estimate the start, end, and inflection points for each PC based on the total number of phrases and the position of each phrase in the utterance. We then draw straight lines between the points within each phrase. To create the ACs, we use a cosine template that is stretched to match the duration of the foot. The AC starts and ends at 0 Hz. Using further simplifications of the LAM, this template is non-uniformly time-warped so that the peak is located at a fixed distance into the foot, depending upon the number of syllables in the foot and the location of the foot in the phrase [9]. Thus, alignment is speaker-independent. The parts of the AC before and after the peak are interpolated on the time axis to fit the time frame of the foot with which they are associated. The peak height is determined by the appropriate parameter (one of parameters 4–6, enumerated at the end of Section 3). The final F_0 contour is created by summing the phrase and accent curves.

4.2. Recombinant Synthesis

To synthesize an F_0 contour from component curves, we require foot and phrase boundary information, as well as the values of the features given in Table 1 for each foot. We first look up PC polynomial coefficients in the corpus, and then ACs, using these feature values. Our corpus does not contain all combinations of feature values. To address this issue, we match the features in the following order, falling back on other feature values if there are no matching component curves: (1) phrasal position, (2) number of syllables/feet in the foot/phrase, (3) sentence type, (4) emphasis type and (5) level of emphasis. syllables/feet in foot/phrase are one, two, or three or more. Finally, if there is more than one component curve in the corpus that matches the input features, we select among them at random.

We first synthesize a PC of the given duration using the polynomial coefficients found in the corpus. We raise or lower the PC after the inflectional point at the beginning of the final foot in the phrase so that there is no discontinuity between the two segments of the PC. Next, we scale the ACs. For each AC, we are given a target length. We rescale the original AC linearly to the target length. Finally, we align the synthetic PC and AC components and add them together to create the synthetic F_0 contour.

Note that our approach to rescaling ACs is extremely simple, and as a result it may not capture interactions between F_0 and segmental, durational or spectral information, for example F_0 peak alignment. However, it is an obvious baseline approach to rescaling ACs. We plan to compare different methods of rescaling ACs in the future.

Stimulus	F_0 Source
RAW	Natural speech
NAT	Delexicalized natural speech
PARAM-SPK	Speaker-specific parametric synthesis
REC	Recombinant synthesis
PARAM-GEN	Gender-specific parametric synthesis

Table 3: Stimulus categories

5. Perceptual Experiment

5.1. Stimuli

The different stimuli in the perceptual experiment are outlined in Table 3. All of the stimuli except for RAW were delexicalized using a low-pass filter with a cutoff frequency of 500 Hz. The synthetic stimuli were produced with the same synthesizer, but different models of F_0 (as described in Table 3). To synthesize speech, we used an in-house unit-selection concatenative synthesizer with prosodic signal modification. To produce the parameter sets for PARAM-GEN, we simply calculated one set of parameters using all of the females' data as input, and another using the males'. For PARAM-SPK, we computed a set of parameters for each speaker using the same algorithm. The stimuli only included declarative sentences. We synthesized the parametric F_0 contours using the method described in Section 4.1, and the recombinant contours using the method in Section 4.2.

5.2. Procedure

We performed a set of listening experiments on Amazon Mechanical Turk (AMT). The experiment was completed by 106 subjects, all of whom had completed at least 100 tasks, had approval ratings of at least 95%, and were located in the USA.

We selected six speakers from the FP, three of each gender, and conducted a block of comparisons for each of them. At the beginning of each block, we presented the listener with a 30s sample of RAW speech from the target speaker. We instructed the listener that they would listen to recordings made through a wall, and that they would tell us whether the person behind the wall is the same as the target speaker. After listening to this introductory sample, the listener was presented with 13 pairs of stimuli. The first stimulus within each pair was a sample of RAW speech from the target speaker, and the second sample was delexicalized speech. Within each block, we presented the following stimuli: 4 $\times$ NAT; 4 $\times$ PARAM-SPK; 4 $\times$ REC and 1 $\times$ PARAM-GEN. For all of the stimuli, aside from PARAM-GEN, we presented two samples from the target speaker, and one sample from each of the other two speakers of the same gender. We did not compare speakers across genders. The order of the blocks was randomized within each experiment, and the stimuli were randomized within each block. Finally, the text of the two utterances was never the same within a comparison.

We asked the listeners whether both sounds were produced by the same speaker, i. e. the target speaker, and they gave their answer on a four point scale: definitely not, probably not, probably yes, and definitely yes. The listeners were shown whether their answer was correct after each question.

To mitigate the risk that subjects from AMT would adopt undesirable strategies (for example, answering that the speakers were the same if the average pitch sounds similar, otherwise different), we rewarded and punished their responses with a cumulative bonus system. Each correct answer added to their cumulative bonus, with each confident answer adding 2¢ and

Method	Males	Females
NAT	0.60	0.68
RECOMB	0.55	0.69
PARAM-SPK	0.55	0.67
PARAM-GEN	0.46	0.61

Table 4: Proportion of correct responses by method and gender of speakers

	Male		Female	
	$t(105)$	p	$t(105)$	p
NAT:P.-SPK	2.26	0.026*	0.24	0.808
NAT:RECOMB	2.18	0.032*	-0.89	0.371
P.-SPK:P.-GEN	2.81	0.003*	2.32	0.011*
RECOMB:P.-GEN	2.77	0.003*	2.81	0.003*

Table 5: Planned t-tests comparing effectiveness of each method

each less-confident answer 1¢. Confident wrong answers reduced their cumulative bonus by 2¢, and less-confident wrong answers by 1¢. If the bonus was negative at the end of the task, we only paid them the base payment, and if it was positive, we paid them the bonus in addition to the base payment. We only rejected a listener's work if it was incomplete. In total, we paid the subjects $133.25, of which $33.25 was bonuses.

6. Results and Discussion

Each subject's response was categorized as correct or incorrect. Answers were considered correct if they correctly identified the speakers as being the same or different, regardless of confidence level. Table 4 shows the proportion of correct responses by method and gender.

We compared the accuracy yielded by different methods by performing paired t-tests in which we paired the percentage of samples from method A correctly identified by a particular subject with the percentage of samples from method B correctly identified by that same listener. Since the percentages of correct responses of male and female speakers were so different, we only compared methods within genders. We compared delexicalized natural speech, NAT, with the two speaker-specific models, PARAM-SPK and RECOMB using two-way t-tests. We compared two speaker-specific methods, PARAM-SPK and RECOMB, to the speaker-independent method, PARAM-GEN with one-way t-tests. Table 5 shows the results of all of these comparisons.

For female speakers, the speaker-specific methods convey speaker ID as effectively as natural speech, and more effectively than the speaker-independent method. With male speakers, we see that while both speaker-specific models convey speaker ID better than the speaker-independent method, they do not do so as effectively as natural F_0.

Finally, we subjected all of these comparisons to a post-hoc Tukey's HSD test, summarized in Table 6. This test confirms that for females, both speaker-specific methods convey speaker ID as effectively as natural speech, and that PARAM-SPK does so better than the speaker-independent method. However, it does not confirm that RECOMB conveys speaker ID as effectively as natural speech. The results of this test suggest that for males, both speaker-specific methods may in fact convey speaker ID as effectively as natural speech, and that both speaker-specific methods may do so more effectively than the speaker-independent method.

7. Conclusion

We have presented a simple method for decomposing F_0 contours into the component curves posited by the LAM. We use these decomposed F_0 contours to train two speaker-specific models of F_0 by estimating parameters, and storing a corpus of F_0 component curves. We use these parameters and component curves in two methods for synthesizing F_0 contours. The parametric method only requires foot boundary information, while the recombinant method also requires features describing the target utterance, for example sentence type and phrasal position. In a perceptual experiment, we have found that both speaker-specific models of F_0 are able to convey speaker identity in the absence of any spectral, durational or phonemic information. Furthermore, we have found that F_0 contours synthesized with either speaker-specific method considered here may convey speaker identity as effectively as natural ones, particularly for female speakers.

Future work should address the limitations of this study. First, we do not know if mean and variance of F_0 are sufficient to convey speaker identity. One could address this issue by modifying LAM parameters and the resulting F_0 contours to have the same mean and variance as the contours produced by PARAM-SPK, and then evaluating these contours in a perceptual experiment like the one described here. Second, we ignore the ways in which F_0 interacts with other information in speech, for example peak alignment and pitch accent location. These interactions may carry speaker-identifying information, but they are destroyed in delexicalized speech, which is all we have considered here.

	Male	Female
NAT:SPK	0.346	0.909
NAT-REC	0.299	0.999
SPK-GEN	0.004*	0.006*
REC-GEN	0.004*	0.057

Table 6: Adjusted p values from Tukey's HSD applied to comparisons in Table 5

8. References

[1] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 4. IEEE, 2003, pp. IV–788.

[2] J. van Santen and B. Möbius, "A quantitative model of F0 generation and alignment," in *Intonation – Analysis, Modelling and Technology*, A. Botinis, Ed. Kluwer academic publishers, 2000, pp. 269–288.

[3] J. Van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3, pp. 365–375, 2005.

[4] E. Klabbers, T. Mishra, and J. van Santen, "Recombinant speech synthesis: Natural text-to-speech synthesis with prosodic control," *Journal of the Acoustical Society of America*, vol. 126, no. 4, p. 2205, 2009.

[5] E. Klabbers, A. Kain, and J. van Santen, "Evaluation of speaker mimic technology for personalizing sgd voices," in *Proceedings Interspeech 2010, Makuhari, Japan*, 2010, pp. p2154–2157.

[6] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," *The production of speech*, pp. 39–55, 1983.

[7] J. van Santen, T. Mishra, and E. Klabbers, "Estimating Phrase Curves in the General Superpositional Intonation Model," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[8] T. Mishra, J. Van Santen, and E. Klabbers, "Decomposition of pitch curves in the general superpositional intonation model," *Speech Prosody, Dresden, Germany*, 2006.

[9] E. Klabbers and J. Santen, "Clustering of foot-based pitch contours in expressive speech," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.