A Statistical Phrase/Accent Model for Intonation Modeling

Gopala Krishna Anumanchipalli†‡, Luís C. Oliveira†, Alan W Black‡

‡Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA 15213
†Spoken Language Systems Laboratory, INESC-ID/IST Lisboa, Portugal 1000-029
{gopalakr, awb}@cs.cmu.edu, lco@inesc-id.pt

10.21437/Interspeech.2011-36

Abstract

This paper proposes a statistical phrase/accent model of voice fundamental frequency(F0) for speech synthesis. It presents an approach for automatic extraction and modeling of phrase and accent phenomena from F0 contours by taking into account their overall trends in the training data. An iterative optimization algorithm is described to extract these components, minimizing the reconstruction error of the F0 contour. This method of modeling local and global components of F0 separately is shown to be better than conventional F0 models used in Statistical Parametric Speech Synthesis (SPSS). Perceptual evaluations confirm that the proposed model is significantly better than baseline SPSS F0 models in 3 prosodically diverse tasks – read speech, radio broadcast speech and audio book speech.

Index Terms: Intonation Modeling, F0, Statistical Parametric Speech Synthesis

1. Introduction

As speech synthesis quality improves, listeners and applications developers have become much more aware of the subtle aspects of how prosody effects the interpretation of synthetic speech. More traditional unit selection [1] speech synthesis techniques have, to a large part, by-passed the issue of explicitly modeling F0 contours by relying on selecting natural contours from a database of natural speech but as we require more control on the F0 for different styles, emotions etc., we must again start to explicitly model F0 contours in order to synthesize the desired range of speech output. In this work, we try to overcome the drawbacks of conventional F0 models in SPSS. We propose a new model that has both the practical flexibility and theoretical basis for improved statistical F0 modeling. The proposed model has two components, one to represent long-term trends (phrases) and the local phenomena (accents). We describe an Expectation-Maximization algorithm to statistically train the components from speech data. The trained model is integrated as an intonational model within a statistical parametric synthesis framework. Intonation contours produced by the proposed approach and those by default F0 models in SPSS are compared both objectively and subjectively on different styles of speech.

2. A Conventional SPSS F0 model

Traditional high quality speech synthesizers, both commercial and research systems, use the unit selection approach or a close variant, where natural chunks of speech (consequently their F0 contours) from natural utterances are pieced together based on contextual information to synthesize novel sentences. Unit Selection has no explicit notion of F0 modeling, and requires a large amount of data from the target domain for optimal approximation of natural prosody. [2] proposes a unit selection approach to F0 contour generation, but still requires large number of instances to have appropriate coverage. While the resulting voices in unit selection approaches are of natural quality within the trained domain, the disadvantages remain– the heavy data requirement, size of the model and inflexibility to new domain or style of speech. To address these, statistical parametric approaches for speech synthesis are gaining much focus.

Statistical Parametric Speech Synthesis involves modeling separate decision trees (CART) for duration, spectra (vocal tract features) and F0 (voice source fundamental frequency) of phoneme states as described in [3]. The intermediate nodes of the decision trees are textual and context questions about the phoneme being synthesized. The leaf nodes are usually Gaussian models, storing the means and variances of the training instances clustered at that node.

In this work we use Clustergen [4], an SPSS system where the F0 is modeled as a continuous contour (interpolated through unvoiced regions of the utterance, except silence regions). In training, F0 values for each frame (over a duration of 5-10 milliseconds) are modelled using a decision tree based on contextual questions. Figure 1 compares the output F0 contour of such a decision tree model to a reference natural contour of a novel sentence. It can be seen that natural F0 has a wider dynamic range and is manifested to convey affective information like emphasis (word prominence). The synthetic F0 contour however lacks any 'interesting' excursions and has a relatively lower overall variance than the natural utterance. This flatness of synthesized contours is often what is perceived by users as being monotonic or 'robotic' in speech synthesis technologies.

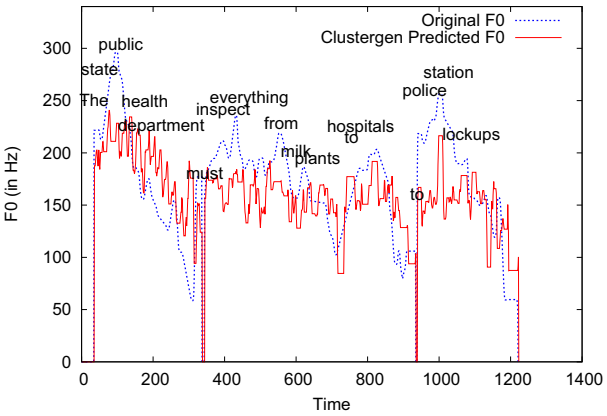

Figure 1: Illustration of original/predicted F0 contour of the sentence *"The state public health department must inspect everything from milk plants to hospitals to police station lockups"*. Words along the original contour show their relative F0 value.

3. Related literature

Intonation is described in several ways in existing literature. We briefly review some that are relevant in the context of this work. A comprehensive survey of intonation models may be found in [5].

3.1. Phonological description

Primarily a representational framework, (Autosegmental) phonological view of the F0 contour is based on the hypothesis that the contour may be realized as a sequence of discrete categories of 'known' shapes that occur as peaks or valleys [6]. Tones and Break Indices, or ToBI [7] is a widely known scheme extending Pierrehumbert's work for annotating English prosody using symbols **H** & **L** to denote different high/low patterns observed in the F0 contour. The framework allows for marking accents and phrase boundaries.

These labels need to be annotated by experts, otherwise there is likely to be disagreement, especially when comparing types of accents.

3.2. Physiological description

Another view of intonation is Fujisaki's production model of F0[8]. Fujisaki formulated a mathematical model that generates the logarithmic F0 contour by addition of three components, the *baseline*, *phrase command* and *accent command*. Baseline is the minimum value of log(F0); phrase and accent, respectively are the long term trend and short term excursion within the contour. In its classical form, the Fujisaki model uses critically damped second order filters to generate the phrase and accent commands for approximating the contours of Japanese declarative sentences. The coefficients of the filters are described to be invariant for a speaker.

The hard assumption within this model is the shape of the components. They are defined to be 'falling'. However, F0 contours of question utterances rise towards the end, making them not directly realizable within this model. There are methods like [9], [10] etc., which build on the Fujisaki model's premise of superimposable components.

3.3. Phonetic Stylization

Among several F0 stylization algorithms is the Tilt Intonation model [11]. The Tilt model provides a continuous description of the F0 contour in terms of parameters that can automatically be derived and synthesized. Within the Tilt framework, the F0 contour is viewed as a series of rise-fall events joined by straight line connections. Each rise-fall event is described by a 4 valued tuple (peak position, amplitude, duration and tilt), which on synthesis gives a perceptually lossless approximation of the event. The model itself does not relate the events to any linguistic unit. The advantage of the tilt representation is that it can concisely represent and synthesize any arbitrary shape of F0 contour. The model has been successfully used as a parameterization for F0 in TTS by [12].

The only requirement is that intonation labels indicating potential regions of rise-fall event are available.

4. Motivation for the proposed model

In this work, we draw upon strengths of existing representations to design an F0 model for SPSS. We use a variant of the Tilt representation where *every* syllable's F0 shape is described as a Tilt 4-tuple (in contrast to only accented syllables being explicitly modeled). This gives complete control over the generated contour to synthesize multi-syllable events and also removes the requirement of intonation labels. Furthermore, we use the tilt model not to represent the actual values of F0 but the 'residual', after appropriately subtracting a phrase component. This is motivated from the Fujisaki model. The notion of independent underlying components, besides being physiologically appealing, also gives an explanation as to why synthetic F0 contours generated by statistical methods (like Figure 1) look 'averaged out'. Previous solutions to this problem included using the global variance of the reference natural data and imposing it on the generated parameters to simulate naturalness [13]. In this work, we take a more theoritical approach, assuming that the fundamental frequency has two underlying additive components. If not separated, these components can nullify each other and corrupt the final model (eg., the down-drift phenomenon reduces the height of the contour in later regions of a phrase, causing two qualitatively equivalent accents in different regions of the phrase to be treated differently). As for the shapes of accents, we use the hypothesis from phonological intonation theories that there are a finite number of 'known' shapes that can describe an excursion in the F0 contour. However, we refrain from predefining their shapes (either of the phrases or the accents), instead letting them be learned from data.

5. Statistical Phrase/Accent F0 model

The proposed model has two components. The same nomenclature *Phrase* and *Accent* is used in reference to these components, like in the Fujisaki model. The F0 values are converted into the log-domain to justify splitting into components. The phrase component is modeled as a CART tree using an appropriate set of long range features. The accents are modeled as a codebook of 'k' speaker specific accent shapes that add to the phrase components to produce the contours. Another CART tree is trained at the syllable level to model which of the k accents is optimal to use given the local context features of the syllable. This codebook CART tree along with the trained k code vectors forms the accent component. The shapes themselves are parameterized by the Tilt representation. Given this model definition, we present an approach to train such a model from data in the next section.

5.1. Constrained Iterative F0 decomposition

We employ an iterative Expectation Maximization algorithm to train the phrase/accent components. An initial estimate of phrase command is used to start the procedure. We use the minimum value of F0 over a syllable as an approximation of the phrase component[1]. Figure 2 illustrates a phrase component initialization as the minimum value of the contour over each syllable.

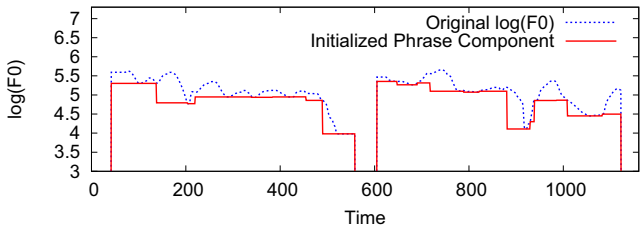

Figure 2: Illustration of Phrase component initialization

[1]Other reasonable initializations lead to similar converged models

For each syllable, the residual (i.e., $log(F0) - phrase$) is parameterized as a 4-valued Tilt tuple. At this stage, to generalize over the training data, the following constraints are applied

- For the phrase components, at each iteration a CART tree is built to regress from long range features, like phrase number, word number within phrase, syllable position in word etc., to the mean value of the phrase at each phoneme (done at the phoneme level for a sharper resolution).
- For the accent components, the constraint is that they should be limited in number. A $k\text{-}means$ clustering is performed to identify the representative shapes of accents over all syllables.

Since the components are trained over the entire training data, they are also robust to utterance specific artifacts of the speaker or pitch detection routines. Also, the constraints are chosen to be minimally assuming and are generic across languages, speakers or speaking styles, giving the model more degrees of freedom. After the intermediate models are built (phrase CART tree and accent codebook), a new estimate $\widehat{log(F0)}$ is reconstructed. The mean reconstruction error over each syllable is added to the previous baseline and residuals are recomputed. This procedure is repeated till an objective criterion is met, here it is the minimum F0 reconstruction error. The parameters that give the best reconstruction error are chosen as the optimal phrase and accent components. A pseudocode of this method is provided as Algorithm 1 below –

Algorithm 1: Constrained Component Extraction

for all utterances **do**
  for all syllables **do**
    set $phrase$ to $min\{F0\}$
    set $accent$ to $tilt(F0 - phrase)$
  end for
end for
while $error \geq \epsilon$ **do**
  train an $accent$ codebook of size k over all accents
  train a $codebook$ CART tree using local features
  train a $phrase$ CART tree using long range features
  for all utterances **do**
    Generate $\hat{F0}$ using $phrase$ & $accent$ codebook
    for all syllables **do**
      accumulate $error$ $(\hat{F0} - F0)$
      update $phrase$ to $(phrase + error)$
      update $accent$ to $tilt(F0 - phrase)$
    end for
  end for
end while

6. Experimental setup

6.1. Speech Databases

We evaluate the proposed model and the training algorithm using several speech databases. We choose 8 speakers from 3 distinct speaking styles. Three sources are used: 2 speakers (`rms, slt`) from ARCTIC [14], a read speech database of short declarative sentences selected from a collection of stories; 5 speakers (`f1a,f2b,f3a,m1b,m2b`) from BURSC [15], a radio broadcast corpus & 1 female speaker's digital audio book (`emma`) of Jane Austen's *Emma* from *librivox.org*. The databases are automatically segmented, aligning the speech with the transcription at a phonetic level. Pitch contours are extracted using the `get_f0` tool of ESPS software [16] and smoothed and interpolated through unvoiced regions to enable modeling F0 as a continuous phenomenon. 8 statistical voices are built, one for each speaker.

6.2. Phrase/Accent Component Training

The iterative F0 decomposition algorithm described in Section 5 is used to extract the phrase and accent components of the F0 contours of all utterances of each speaker. Table 1 shows some features used in training the component models.

Phrase features (Global trend)	Accent features (Local excursion)
word POS	word POS
phrase number	syllable category
word position in phrase	predicted accent
#syllables in phrase	lexical stress
content words in phrase	prev/next values
normalized values of above	

Table 1: Example features used to train Phrase/Accent models

Note that conventional F0 CART models use all features together in the model training. But in the proposed method, we separate them to appropriately deal with the phrase and accent components separately. Table 2 presents a trace of the training algorithm on one speaker `f1a`. It can be seen that the overall root mean squared error(RMSE) decreases and correlation(CORR) increases on the training data before converging over iterations.

#Iter	1	2	3	4	5	6	7	8
RMSE	0.457	0.384	0.237	0.186	0.182	0.181	0.181	0.180
CORR	0.488	0.549	0.641	0.705	0.714	0.717	0.718	0.719

Table 2: RMSE/Correlations per training iteration on task `f1a`

The best average resynthesis error is about 1-1.2 Hz for all speakers, which is perceptually insignificant. An example best component split is shown in Figure 3 where the F0 is plotted along with the derived phrase and accent components and the resynthesized contours. It can be seen that the final derived phrase is a gradual falling contour (though no such constraint is explicitly enforced in the model) and the accents are sequences of what look like metrical feet spread over multiple syllables.

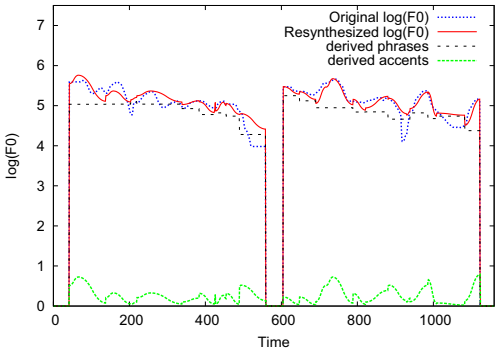

Figure 3: Example F0 contour split into best phrase and accent commands

The trained phrase and accent components can be used as an intonation model and used for synthesizing F0s of novel sentences. At test time, the phrase and trees are traversed to predict the best possible long-term trend curve and local excursion

sequence and added to generate a contour for a novel sentence. Figure 4 compares the predicted contours of an unseen sentence generated by default F0 model in Clustergen and the proposed statistical Phrase/Accent model. It is easy to see that the F0 generated by the proposed approach has better variance and is seemingly more affective than the default Clustergen F0 model.

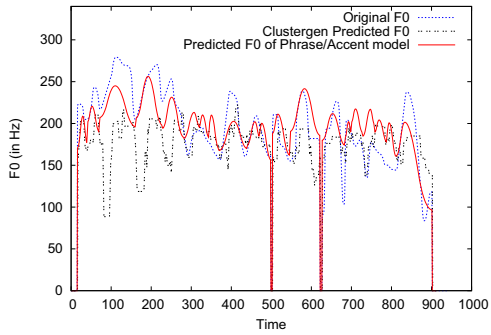

Figure 4: Predicted F0s of Phrase/Accent vs Clustergen model

6.3. Evaluation

Table 3 objectively compares the contours generated by Clustergen F0 model and the proposed statistical Phrase/Accent model on all 8 voices in terms of mean error and correlation. The proposed model scores comparably, yet worse than the default model in most cases. It is to be noted, however that RMSE and CORR measures are computed at the frame level (5-10ms) which may be unsuitable for comparison of intonation contours that have a much higher resolution. This observation is in keeping with earlier studies showing that these measures for comparison of synthetic F0 contours may not be ideal[17].

	Clustergen		Phrase/Accent	
Task	RMSE	CORR	RMSE	CORR
`rms`	10.50	0.66	13.52	0.55
`slt`	11.15	0.63	14.25	0.55
`f1a`	29.85	0.44	30.79	0.55
`m1b`	14.80	0.45	17.05	0.40
`f2b`	28.23	0.57	29.96	0.54
`m2b`	23.49	0.42	25.65	0.37
`f3a`	27.83	0.35	30.49	0.55
`emma`	41.58	0.09	45.11	0.15

Table 3: Objective comparison of voices

To get a more reliable comparison of the two approaches, we conduct subjective AB listening tests, where human listeners are presented with a pair of speech stimuli, same in all respects except the intonation. 3 tasks `slt`, `f2b` and `emma` from each speaking style are used for the listening tests. 10 unseen sentences from each task are synthesized using the default and proposed F0 models. 11 American English speakers were presented the stimuli in a random order and asked to judge which sample they prefer. Fig 5 summarizes the user responses. It can be seen that the proposed Phrase/Accent approach is preferred by listeners in over 80% of cases conclusively showing that the proposed model generates more natural intonation contours than the default model, irrespective of the speaking style.

7. Conclusion

In this paper, we present a statistical phrase/accent model of F0 for text-to-speech within the paradigm of statistical parametric synthesis. An expectation maximization like algorithm is presented to automatically learn the model components from data. The 2-component representation is used as an F0 model in a real TTS system and trained to generate speech examples with predicted contours. The intonation contours thus generated are adjudged by human listeners to be significantly more acceptable than those of conventional methods.

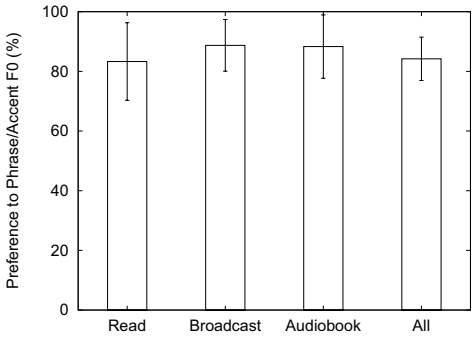

Figure 5: Subjective preference to proposed model over Clustergen (with 95% confidence interval)

8. Acknowledgements

This work was supported partly by the Fundação de Ciência e Tecnologia through the CMU/Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.

9. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP-96*, vol. 1, May 1996, pp. 373 –376 vol. 1.

[2] A. Raux and A. Black, "A unit selection approach to F0 modeling and its application to emphasis," in *Proc. ASRU 2003*, US Virgin Islands, 2003.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] A. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Interspeech 2006*, Pittsburgh, PA., 2006.

[5] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[6] J. B. Pierrehumbert, "The phonology and phonetics of english intonation," *Thesis (Ph.D.)–MIT, Dept. of Linguistics and Philosophy*, 1980.

[7] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, P. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *ICSLP 1992*, Alberta, Canada, 1992.

[8] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, P. MacNeilage, Ed., 1983.

[9] G. Bailley and B. Holm, "SFC: A trainable prosodic model," *Speech Communication*, vol. 46, 2005.

[10] J. van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, pp. 365 – 375, 2005.

[11] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107 3, pp. 1697–1714, 2000.

[12] K. Dusterhoff, A. Black, and P. Taylor, "Using decision trees within the tilt intonation model to predict F0 contours," in *in Proc. Eurospeech 1999*, 1999, pp. 1627–1630.

[13] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE - Trans. Inf. Syst.*, vol. E90-D, pp. 816–824, May 2007.

[14] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, CMU, PA, Tech. Rep., 2003.

[15] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "Boston university radio speech corpus," no. LDC96S36, 1996.

[16] "Entropic signal processing system (ESPS)," Entropic Inc., 1999.

[17] R. Clark and K. Dusterhoff, "Objective methods for evaluating synthetic intonation," in *Proc. Eurospeech 1999*.