

F0 RANGE AND PEAK ALIGNMENT ACROSS SPEAKERS AND EMOTIONS

Eric Morley*

Jan van Santen*[†]

Esther Klabbers*[†]

Alexander Kain*[†]

*Center for Spoken Language Understanding
Oregon Health & Science University
Portland, OR 97239, USA

[†]BioSpeech, Inc.
940 Upper Devon Lane
Lake Oswego, OR 97034, USA

ABSTRACT

We present an analysis of F_0 range and peak alignment in emotional speech from a heterogeneous group of speakers varying in age and gender. Both speaker and emotion had a strong effect on F_0 range. Despite these large changes in the F_0 trajectory, peak alignment was remarkably stable. Using the Linear Alignment Model (LAM) [1], we show that the effects on alignment of emotion and speaker differences, although statistically significant, are small. This stability results in a conclusion that peak alignment, unlike F_0 range, does not appear to carry much information about speaker identity or emotional state. The LAM is effective in that it explains 42% of the variance in peak location on average, and furthermore it predicts the time of F_0 peaks with an average RMS error of 12ms.

Index Terms— speech analysis, speech synthesis, human voice, emotion recognition

1. INTRODUCTION

Previous studies have found that F_0 range and mean varies significantly between emotions (e. g. [2, 3]). In this paper, we investigate whether in addition to the amplitude, the time alignment of F_0 peaks differs between speakers, across emotions, or both. This knowledge is important for designing personalized voices for a synthetic speech system: if F_0 peak alignment varies across either speakers or emotions, then the intonation model should reflect this (unless further studies find that such differences are imperceptible); if there are no such differences, then there is no need to change the parameters governing F_0 peak alignment when constructing personalized voices.

2. THE LINEAR ALIGNMENT MODEL

The Linear Alignment Model (LAM) provides a way to quantify and compare pitch alignment for sets of utterances, for

example utterances spoken with different emotions or by particular speakers [1]. This model assumes that utterances are divided into stress-timed feet, which are defined as an accented syllable followed by any number of unaccented syllables without regard to word boundaries [4].

The LAM allows one to predict the temporal position of various points on the F_0 trajectory. In this investigation, however, we are only interested in the time of the F_0 peak of the stressed syllable (T_{peak}), as measured from the start of the foot. The LAM claims that T_{peak} is given by the following formula:

$$T_{\text{peak}} = \alpha D_{\text{onset}} + \beta D_{\text{s-rhyme}} + \gamma D_{\text{rest}} \quad (1)$$

In the equation above, D_{onset} is the duration of the onset of the accented syllable, excluding any non-syllable-initial sonorants. $D_{\text{s-rhyme}}$ is the total duration of the s-rhyme, which is the nucleus, and any sonorants in the accented syllable, excluding syllable initial sonorants. D_{rest} is the duration of the rest of the foot. These three D variables will be referred to collectively as sub-foot durations.

Most formulations of the Linear Alignment Model use multiple sets of phonetically-sensitive α, β, γ parameters. One formulation estimates three sets of parameters, one for each of three consonant classes (voiceless, voiced non-sonorant, voiceless). The consonant class of all three parameters is determined by the consonant in the onset of the foot [5]. Another formulation calculates nine sets of parameters. In this formulation the choice of parameter set is determined by both the onset consonant of the foot, and the consonant immediately following the s-rhyme [6]. The consonant classes in the second formulation are identical to those in the first.

In this investigation, we calculate only one set of parameters, which is used for all consonant classes. This is primarily due to data sparsity; we simply do not have enough data to reliably estimate three or nine sets of parameters for each speaker and for every emotion.

3. EXPERIMENT

3.1. Speech Corpus

The utterances used in this experiment are part of the CSLU Emotion corpus [7]. The corpus contains 24 different utter-

This research was supported by NIH grants R01-DC007129, R21-DC010239, and R42-DC008712, and by NSF grants IIS-0205731 and IIS-0964468. The views herein are those of the authors and do not reflect the views of the funding agencies.

ances spoken with each of four emotions: anger, fear, happiness and sadness. To elicit these utterances, 24 vignettes were written for each of the four emotions, such that these vignettes could be grouped by fours that all ended in the same “target” sentence. It is these “target” sentences which make up the corpus. This experiment used data from 12 different speakers (8 female adults, 3 female children, 1 male adult).

In terms of ToBi labels, all of the stressed syllables which begin each foot in these data are associated with standard non-contrastive H* pitch accents. Contrastive stress and diverse intonational patterns can therefore be ruled out as a contributing factor to any variance in peak alignment.

The data were annotated by hand before analysis. First, sub-feet (onset, s-rhyme and rest as described above) were labeled. F_0 values were calculated every 10 ms for each utterance using the ESPS method and F_0 peaks were labeled by hand.

3.2. Estimating Parameters α, β, γ

After annotating the data, we estimated α, β, γ parameters for each speaker/emotion pair p . This was accomplished using the following procedure: for every foot f in the data for p : (1) perform robust linear regression on all feet except for f , using the sub-foot durations as predictor variables and the T_{peak} as the indicator variable; (2) record the estimated α, β, γ parameters in row f of the matrix M . The final values of α, β , and γ for p are the median values of each column of M . Briefly, robust linear regression uses the iteratively reweighted least-squares method to diminish the influence of outliers on the estimated parameters [8].

Van Santen et al. found that $1 > \alpha > \beta > \gamma > 0$ in neutral speech [5]. This does not appear to be the case with emotional speech. In all cases $\alpha > \beta, \gamma > 0$, but there are cases where $\alpha > 1$ and where $\gamma \geq \beta$. In the interest of space, a summary of the estimated parameters for all speaker/emotion pairs, rather than all of the estimates, are given in Table 1. Histograms for each of the parameters (with the parameter value along the x-axis and the number of speakers along the y-axis) are shown in Figure 1. Note that the Lilliefors test shows that all of the parameters, with the exception of the α and γ parameters for the sad affect, are normally distributed.

3.3. Predictive Quality of Estimated Parameters

We used the estimated α, β, γ parameters and the manually determined sub-foot durations to predict T_{peak} for all of the feet in the data. The predicted values of T_{peak} were then compared with the true peak times.

Two measures of the effectiveness of this model, and the predictive power of the estimated parameters, are *root mean square error* (RMS error) and *correlation*. In this context, RMS error quantifies how close the predicted values of T_{peak} are to the actual values. Correlation measures how much of the variation in F_0 peak placement is explained by the LAM.

	Angry			Fearful		
	α	β	γ	α	β	γ
Mean	0.88	0.29	0.19	0.83	0.24	0.26
Std.	0.14	0.10	0.06	0.21	0.08	0.06
Max.	1.15	0.42	0.29	1.35	0.38	0.36
Min.	0.70	0.09	0.10	0.61	0.12	0.16
	Happy			Sad		
	α	β	γ	α	β	γ
Mean	0.77	0.33	0.26	0.77	0.26	0.22
Std.	0.18	0.05	0.06	0.18	0.10	0.08
Max.	1.07	0.42	0.38	0.94	0.54	0.33
Min.	0.52	0.25	0.17	0.27	0.16	0.07

Table 1: Summary of estimates of α, β, γ

The mean RMS error over all of the predicted values is 12 ms, and the mean standard deviation is 8 ms. These statistics are also identical within each speaker/emotion pair. The overall maximum RMS error is 35 ms, and the minimum is < 1 ms. Given that F_0 is calculated at 10 ms intervals, the RMS error is low.

The LAM explains a high amount of the variance in F_0 peak placement as well. The median correlation between the predicted and actual values of T_{peak} is 0.66, and the mean value is 0.65. This means that on average, LAM explains approximately 42% of the variance in F_0 peak placement within each speaker/emotion pair. Noise within this experiment, for example in F_0 calculations and F_0 peak labeling, are both sources of potential variance in the measurements. Using a single set of parameters for all utterances, rather than a different set for each consonant class is another potential source of variance; higher correlation would be expected with context-specific parameters.

4. PARAMETER ANALYSIS

4.1. Differences between Emotions and Speakers

The E-statistic (energy) test of multivariate normality was applied to the estimated parameters α, β and γ [9]. The null hypothesis of the E-statistic test is $X \sim \mathcal{N}(\mu, \sigma^2)$. Here $E = 0.69$, and $p = 0.92$, and therefore the null hypothesis can not be rejected. The data is taken to be normally distributed, thus lending itself to analysis with MANOVA.

The MANOVA model used here is $X \sim (e + s)$ where $e = \text{emotion}$ and $s = \text{speaker}$. This analysis showed that both speaker and emotion have a statistically significant effect on the α, β, γ parameters (emotion: $p = 0.02$, speaker: $p = 0.01$). In quantitative terms, $\mu_{p_i} \neq \mu_{p_j}$ where $p \in \alpha, \beta, \gamma$ and i, j are speaker/emotion pairs which differ in at least one respect.

ANOVA permits a finer grained analysis than MANOVA; specifically, it can show which parameters vary between speakers, and which vary across emotion. The results of the ANOVA performed on each of the three parameters are in

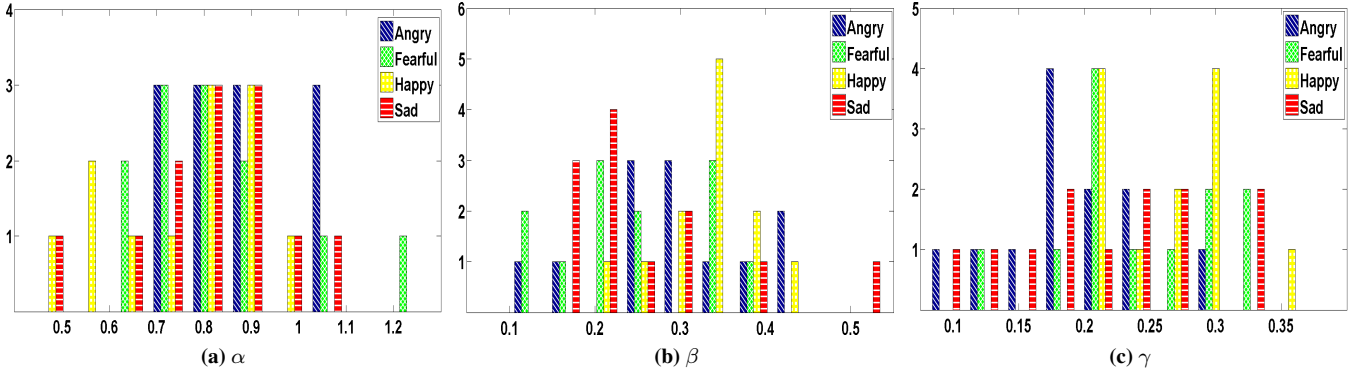


Fig. 1: Histograms of parameter estimates (count along y-axis, parameter value along x-axis)

		Df	SS	MS	F	$\text{Pr}(> F)$
α	Emotion	3	0.098	0.033	1.197	0.326
	Speaker	11	0.510	0.046	1.702	0.116
	Residuals	33	0.899	0.027		
β	Emotion	3	0.060	0.020	2.972	0.046
	Speaker	11	0.107	0.010	1.442	0.201
	Residuals	33	0.223	0.006		
γ	Emotion	3	0.036	0.012	3.300	0.032
	Speaker	11	0.069	0.006	1.701	0.117
	Residuals	33	0.121	0.004		

Table 2: Summary of ANOVA of α, β, γ

Table 2.

These analyses show that both the means of the β and γ parameters vary across emotions. The sizes of these effects, however, are small: partial η^2 for the effect of emotion on β is 0.21, and for γ it is 0.23. Also note that although MANOVA finds a difference in the means of the parameters as a whole between speakers, no individual parameter appears to be responsible for this difference.

4.2. Speaker Identification

Having found that the α, β, γ parameters vary across emotions and speakers, we now consider whether these parameters vary enough across speakers to be used as a speaker identifier within a given emotion. To test whether the parameters indeed can be used in this way, we repeatedly: (1) partition the relevant data randomly into two equally-sized sets; (2) estimate the α, β, γ parameters for each of the speakers in the two sets separately using robust linear regression with the sub-foot durations as predictor values and T_{peak} as the indicator values; (3) store the six parameter estimates in columns 1 and 4 (α), 2 and 5 (β), and 3 and 6 (γ) of a matrix; (4) calculate and record the correlations of each of these pairs of columns (for example columns 1 and 4), which gives a single correlation value for each parameter/emotion pair. We repeated this procedure 10,000 times for each emotion.

The values of interest to us are the mean parameter self-correlation coefficients for each parameter/emotion pair. High

	α	β	γ
Angry	-0.20	0.29	-0.13
Fearful	<0.01	0.09	-0.14
Happy	-0.02	-0.05	0.09
Sad	0.07	0.31	0.09

Table 3: Mean parameter self-correlation coefficients

mean correlation between estimates of a given parameter show that within that emotion, speakers systematically differ on this parameter. Similarly, low mean correlation between estimates of a given parameter show that while the parameter varies between speakers, it does not do so systematically, and therefore it can not be a speaker identifier. This is the case because the parameter estimates should be close regardless of what training data is used.

The mean parameter self-correlation coefficients calculated by this procedure are given in Table 3. From these correlation coefficients, we may tentatively conclude that the α, β, γ parameters cannot be used as speaker identifiers within an emotion.

A potential objection to this procedure is that there may not be enough training data for each emotion to obtain reliable estimates of α, β , and γ . Addressing this concern is quite simple. To do so, we modify the original procedure for calculating parameter self-correlation coefficients as follows: instead of considering emotions separately, we pool all data for a speaker across emotion. As before, this data is then partitioned into two groups of equal size, parameters are estimated and the correlation calculated. This quadruples the amount of data used for each speaker, and should give more reliable estimates. The resulting correlations are α : 0.28, β : 0.28 and γ : 0.05. This shows that while there may be some speaker identification information in the α, β, γ parameters due to their cross-speaker variation, the level is minimal.

5. F_0 RANGE

The above analysis shows that although speakers and emotions have some effect on F_0 peak alignment, their effect sizes are so small that F_0 peak alignment is quite stable between speakers and across emotions. One possibility, which would

	Angry	Fearful	Happy	Sad
Mean	2.43 lgHz	2.39 lgHz	2.47 lgHz	2.34 lgHz
Range	2.46 lgHz	2.27 lgHz	2.48 lgHz	2.19 lgHz

Table 4: Across-speaker means of F_0 mean and range

nullify the finding that it is stable across emotions if true, is that the actors did not actually modify their speech when speaking with different emotions. For that to be the case, one would expect both the fine temporal alignment of the pitch peak and the crude measure of the range of F_0 to vary minimally across both speakers and emotions. Mean F_0 should also vary minimally across emotions if this is true. The mean and range of F_0 will be calculated in $\log_{10}\text{Hz}$ (lgHz), and only in voiced regions.

There are clear differences in both mean F_0 and the range of F_0 across emotions and between speakers. Means of these data (across speakers) are given in Table 4. All of these data, except for the mean F_0 of angry speech, are found to be normally distributed using the Lilliefors test. ANOVA finds that the effect of both emotion and speaker on mean F_0 and F_0 range are highly significant ($p \leq 0.0001$ in all cases). The partial η^2 statistic of 0.57 demonstrates that emotion has a medium-sized effect on mean F_0 across emotions. The data contains utterances produced by men, women and children, and therefore it is unsurprising that speaker differences have a large effect on mean F_0 (partial $\eta^2 = 0.83$).

Speaker differences have a medium-sized effect on F_0 range: partial η^2 for this effect is 0.65. Interestingly, the effect size of emotion is even bigger, as shown by its partial η^2 statistic of 0.73. This confirms that speakers varied their speech when producing different emotions, and that they did not vary it uniformly.

The differences in mean F_0 and F_0 range across speakers and emotions show that there are real differences in the production of each emotion. Furthermore, the effects of emotion on the mean and range of F_0 are largely in the same direction as previous research using acted emotions. Both happy and angry speech showing a higher mean F_0 than sad speech, although previous research has found fearful speech to have a higher mean F_0 than angry speech. The ranges of F_0 in the data analyzed here correspond even more nicely with previous research, which has found that happy and angry speech have a broader range in F_0 than do fearful and sad speech F_0 [2]. In spite of these major differences in F_0 amplitude across speakers and emotions, an aspect of the fine temporal structure of the F_0 trajectory, namely the location of the peak, remains relatively stable in these different contexts.

6. CONCLUSION

The Linear Alignment Model provides a simple, yet effective way of estimating the temporal position of F_0 peaks. This model depends on three parameters, which have been shown to be relatively stable between speakers and across

emotions. In particular, the amplitudes of the F_0 trajectory in the data vary widely as a result of both speaker and emotion, while differences in the parameters governing peak-alignment change only a small amount from these same effects.

There are two obvious practical applications of these findings. The first of these is that even though F_0 peak placement varies significantly across emotions, it does not vary enough to be a useful feature in emotion recognition. The second application is in constructing personalized voices for a synthetic speech system: rather than determining speaker- or emotion-specific parameters for the Linear Alignment Model when it is used to predict the time point of F_0 peak, one can instead use a single set of parameters for all speakers and emotions. This simplifies the task of creating personalized intonational models.

7. REFERENCES

- [1] J.P.H. van Santen and B. Möbius, “A quantitative model of F0 generation and alignment,” in *Intonation – Analysis, Modelling and Technology*, Antonis Botinis, Ed., pp. 269–288. Kluwer academic publishers, 2000.
- [2] A. Paeschke, M. Kienast, and W.F. Sendlmeier, “F0-contours in emotional speech,” in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 1999, vol. 2, pp. 929–932.
- [3] I.R. Murray and J.L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” *The Journal of the Acoustical Society of America*, vol. 93, pp. 1097–1108, 1993.
- [4] D. Abercrombie, *Elements of general phonetics*, Edinburgh University Press, Edinburgh, 1967.
- [5] J. van Santen, E. Klabbbers, and T. Mishra, “Toward measurement of pitch alignment,” *Italian Journal of Linguistics*, vol. 18, no. 1, pp. 161–188, 2006.
- [6] R.W. Sproat, *Multilingual text-to-speech synthesis: the Bell Labs approach*, Kluwer academic publishers, 1998.
- [7] E. Klabbbers, T. Mishra, and J. van Santen, “Analysis of affective speech recordings using the superpositional intonation model,” in *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007, pp. 339–344.
- [8] P.W. Holland and R.E. Welsch, “Robust regression using iteratively reweighted least-squares,” *Communications in Statistics-Theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.
- [9] Gábor J. Székely and Maria L. Rizzo, “A new test for multivariate normality,” *Journal of Multivariate Analysis*, vol. 93, no. 1, pp. 58–80, 2005.