TOWARDS A MODEL FOR THE HUMANISATION OF PITCH DRIFT IN SINGING VOICE SYNTHESIS

Ryan Stables, Dr. Cham Athwal
School of Digital Media Technology, Birmingham City University
Millennium Point, Birmingham, UK
ryan.stables@bcu.ac.uk
cham.athwal@bcu.ac.uk

Dr. Jamie Bullock
Birmingham Conservatoire, Birmingham City University, Paradise Place, Birmingham, UK
jamie.bullock@bcu.ac.uk

ABSTRACT

A model is presented for the analysis and synthesis of low frequency human-like pitch deviation, as a replacement for existing modulation techniques in singing voice synthesis systems. Fundamental Frequency \( f_0 \) measurements are taken from vocalists producing a selected range of utterances without vibrato and trends in the data are observed. Finally, we present a function that provides natural sounding low frequency \( f_0 \) modulation to synthesized singing voices.

1. INTRODUCTION

Singing voice synthesis systems such as [4] and [5] are being used extensively to generate new expressive musical instruments and are becoming increasingly more intelligible, most of which are close to that of human speakers, focus therefore is often on the perceived naturalness of the vocal. In this study we aim to emulate the low frequency modulation in \( f_0 \), in an attempt to remove the perceptually robotic qualities of synthesized singing voices.

In singing, drift is the low frequency, involuntary perturbation of the \( f_0 \) during phonation that occurs below the vibrato frequencies, which in western music is around 5-7Hz. Physiologically, Orlikoff and Bakenboth [3] suggest that laryngeal muscles and the beating of heart are influential in the stability of a singer’s \( f_0 \) contour. Whereas, Burnett [4] suggests that auditory feedback also has a psychoacoustic influence on the mechanism of \( f_0 \) control.

Drift in singing voice synthesis is acknowledged in numerous papers, however most authors approach this as a minor addition to the synthesis procedure. Studies such as Macon’s [5] stress the requirements for human error in the \( f_0 \) contour. This is reflected in the perceptual experiments undertaken by Saitou [6], in which the least natural sounding synthesized singing voices are those with a smoothed \( f_0 \) contour. This smoothing removes the majority of involuntary pitch deviation, demonstrating the importance of drift to human perception.

Alternative methods for generating low frequency modulation, used by both Lai [7] and Macon [5] are derived from a function used in Klatt’s KLGLOT88 vocal synthesizer [8]. The study describes the technique as a quasirandom drift in the fundamental frequency. The original equation uses the sum of three sinusoids as a multiplier for a fundamental frequency. The three chosen frequencies \([4510, 0141, 710]\) allow long periods until repetition. The range of perturbation or amplitude of the deviation is influenced by a flutter ‘FL’ coefficient.

$$\Delta f_0 = (FL/50)(f_0) \left[ \sin(2\pi 12.7t) \right. + \sin(2\pi 7.1t) + \sin(2\pi 4.7t) \right]$$

Klatt states that whilst this approach seems to be sufficient for their singing synthesis system, it is highly unlikely to accurately represent the human deviations from a fixed frequency. The same function is used by Macon in the concatenative synthesis engine developed in 1997, and later by Lai in 2007. Both of the more recent systems use an FL coefficient of 0.33.

2. METHODOLOGY

In this study we aim to estimate the manner in which the features of drift change in response to variation in the range of parameters present in the singing voice. This will allows us to develop a parametric model in which the attempted \( f_0 \), the linguistic content and the level of singing experience effects the \( f_0 \) contour.

The parameters were chosen due to their importance to singing styles in popular music. To measure their influence on the signal’s features, we have compiled a dataset of 10 professional and 25 amateur singers, producing utterances that meet the proposed criteria. Subjects were
asked to listen to a range of sine waves between 100Hz and 800Hz at 70dB through headphones and simultaneously produce an /a/ phoneme without vibrato at the same perceived pitch. Between each sound 5 seconds of silence was played and subjects were able to monitor their own voice through headphones. The process was then repeated for /o/ and /e/ phonemes.

Recordings were taken using a MacBook Pro with a Mackie Onyx sound card and an AKG-414 condenser microphone and analysed using a subharmonic-to-harmonic ratio $f_0$ tracking algorithm [9] with a frame size of 40ms, taken every 10ms. The $f_0$ tracker was chosen due to its accuracy with voiced signals, parameters were set based on the level of resolution required for our experiment.

3. RESULTS

It is evident from the dataset that the recorded $f_0$ contours contain small gaussian-like regions modulated by a low frequency component, this corresponds with the work carried out in [10] and can be seen in figure 1. The figure shows signals with 2 gaussian regions, being produced at 138Hz and 185Hz respectively. The vocalists drift from one Gaussian range to another, with clear shifts in mean ($\mu$) and standard deviation ($\sigma$).

As the attempted $f_0$ parameter increases, the $\mu$ and $\sigma^2$ of the drift signals were measured in order to observe any trends. The results shows that generally, higher $\sigma^2$ values are found at higher $f_0$’s, however there is no clear correlation between the two. The small Gaussian regions in the suggested model are also relatively inconsistent, with no clear trends in the amount, lengths, $\mu$’s and $\sigma$’s.

In order to evaluate the effects of training and singing experience, the measurements from the professional vocalists are separated from the amateur recordings. Figure 2 shows a comparison between typical distributions of the two sources. The dataset shows that singers with classical training generally have a lower $\sigma^2$ and more accurate $\mu$, demonstrating their ability to maintain a lower pitch deviation. The Professional samples also contain fewer regions, and effectively appear more Gaussian.

Three voiced phonemes were measured over a range of frequencies in order to assess the impact of linguistic content on the parameters of drift. From our recordings, the /o/ phoneme frequently had the lowest $\sigma^2$, however features in the model such as number of regions ($R_n$) and mean had no obvious correlation with linguistic content. Table 1 shows the average values from the dataset taken across different frequencies and phonemes.
4. MODELLING DRIFT

To synthesize utterances with small regions of quasi-Gaussian drift, separate by a low frequency modulator. The following method is proposed:

\[ \Delta f_0 = x_i \sigma_i^2 + \mu_i \quad (2) \]

where:

\[ \begin{align*}
  t \geq R_1, & \quad i = 1 \\
  t \geq R_2, & \quad i = 2 \\
  t \geq R_n, & \quad i = n
\end{align*} \]

Where \( x_i \), \( R \), \( \sigma_i^2 \) and \( \mu_i \) contain independent random variables, in order to produce a drift contour with \( n \) regions. This provides a system for mapping pseudorandom normally distributed values to an empirical distribution, based around the observed trajectory of the \( f_0 \) in human contours. The distribution of each region is considered to be Gaussian for synthesis purposes, a result of this function is shown in Figure 3.

![Figure 3. Distributions created by the suggested method.](image)

In order to add parameters such as \( f_0 \), linguistic content and singing experience to the model, a simple Markov Chain is used in order to dynamically weight the \( R \), \( \sigma_i^2 \) and \( \mu_i \) features in Equation 2. For each of the chosen parameters, a State Transition Matrix (STM) is created, and populated with the corresponding feature vectors. This allows us to weight the features with probability coefficients taken empirically from our dataset.

To generate states, both the parameters and the feature vectors are broken down into groups of eight equal bands and paired based on the dataset. The matrices are then used as weightings for the variables in Equation 2. The result of this process is a \( \Delta f_0 \) that contains variations in the amount (n), \( \sigma_i^2 \), \( \mu \) and length of smaller regions, which are relative to the variations in model parameters.

The proposed method is derived from drift that occurs as a result of auditory feedback. A possible explanation for this is that as vocalists hear excessive deviation from a desired frequency in the perceived pitch of their voice, an unconscious process alters this pitch in an attempt to become closer to the target frequency. This automatic process produces small jumps in \( f_0 \) between relatively stable but slowly deviating regions. When the sample points from a human \( f_0 \) contour are mapped to a histogram, smaller clusters of sample points at varying offsets from a mean frequency are evident.

For the proposed modelling technique, we consider these clusters to be a combination of smaller, potentially overlapping, sliding Gaussian curves, occurring at discrete time intervals with regions of relatively low power in between. The individual curves represent regions in the contour at which the \( f_0 \) is relatively stable, the mean frequency for each region in \( x_i \) is an offset value from the global mean, chosen by the Markov chain. Each corresponding standard deviation (\( \sigma_i^2 \)) is limited to \( 1/N \) of the global \( \sigma^2 \), with \( N \) representing the number of regions. A comparison between this technique and the previously defined methods is shown in Figure 4.

5. CONCLUSION

In order to produce a drift signal, which mirrors that of the human \( f_0 \) control system, we have applied a novel probability based function to low-pass filtered Gaussian noise.
We are left with a signal that fluctuates between regions of normal distribution with smaller standard deviations and amplitudes.

This could be attributed to the influence of the auditory feedback mechanism on \( f_0 \) in singing, encountered when a vocalist attempts to match a target frequency. As the subject’s phonation frequency drifts further away from the target frequency, larger jumps in \( f_0 \) are made in an unconscious attempt to become closer to the target.

As trends in the number of regions and the region lengths were not immediately evident over a series of different fundamental frequencies and phonemes, we use a probability model to produce a statistical representation of our dataset. From preliminary listening tests, we area able to produce singing voices which have a more natural, human like deviation in pitch when compared to existing systems.

6. FUTURE WORK

In future work, extensive listening tests are planned. This will evaluate the perceptual relevance the model. In order to do this, subjects will be asked to give a quantitative evaluation of a dataset created using the previous and current techniques. Each stimulus will be given a rating of naturalness, and compared with a human utterance.

Currently, the features in the model are estimated over a small range of parameters. In order to create a robust, comprehensive model, more linguistic components are needed over a much wider range of attempted \( f_0 \)’s. Furthermore, a much wider range of subjects is required, with a variation in singing experience. Other aspects such as vibrato and musical context also need to be included in the study.

7. REFERENCES


