THE HUMANISATION OF STOCHASTIC PROCESSES FOR THE MODELLING OF F0 DRIFT IN SINGING

Ryan Stables[1], Dr. Jamie Bullock[2], Dr. Cham Athwal[3]

[1] Institute of Digital Experience, Birmingham City University, Birmingham, UK
[2] Birmingham Conservatoire, Birmingham City University, Birmingham, UK
[3] School of Digital Media Technology, Birmingham City University, Birmingham, UK
{ryan.stables, Jamie.bullock, cham.athwal} @bcu.ac.uk

Abstract

We present a model for the generation of low frequency human-like pitch deviation. We take f_0 measurements from vocalists producing a 300Hz fixed tone without vibrato and find that smaller regions are evident, each with Quasi- Gaussian distributions. We present a function to implement this with a PSOLA pitch shifting algorithm, providing natural sounding enhancements to singing voice synthesis systems.

Key Words: Fundamental Frequency, Singing Synthesis, Drift, Humanisation

1 Background

In the following study, Fundamental Frequency f_0 control in the singing voice is investigated and a novel technique for the stochastic production of drift is suggested. Drift is the low frequency, involuntary modulation of the f_0 during phonation. In singing this correlates with non- periodic fluctuations occurring below 5hz, which in western music, sit just below the frequency range of vibrato. Drift has been linked closely with physiological and psychological aspects of speech production. Physiologically, Orlikoff and Bakenboth [1] suggest that laryngeal muscles and the beating of heart are influential in the stability of a singer's f_0 contour. Whereas, Burnett [2] suggests that auditory feedback also has a psychoacoustic influence on the mechanism of f_0 control.

At present, singing voice synthesis and transformation systems are being used extensively to generate new expressive musical instruments and effects [3][4]. We have found that by adding low frequency modulation based on a human process, voices are less robotic and closer to that of an actual singer. Humanisation of this modulation technique ultimately provides enhancements to the field of singing synthesis by contributing to more natural sounding voices.

1.1 Previous methods

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. In order to re-synthesize this, a low pass filter is applied to a noise signal with Gaussian distribution. This is added to the waveform at the end of the process.

Alternate methods for generating the 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 quasi-random 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 [12.7, 7.1, 4.7] allow long periods until repetition. The range of perturbation or amplitude of the deviation is influenced by a flutter 'FL' coefficient.

$$\Delta f_0 = (\mathrm{F\ L}/50)(\mathrm{f0})[\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t)] \quad (1)$$

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 a 0.33' amplitude scaling coefficient (FL/50).

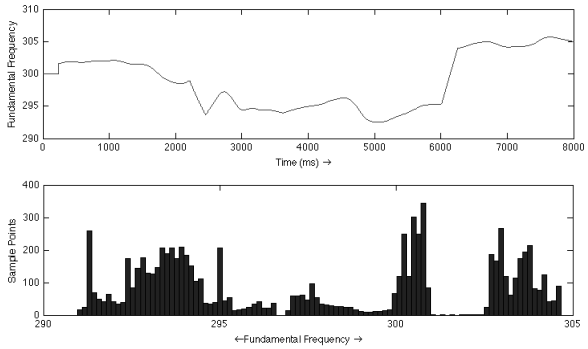

Figure 1: Human f_0 contour and corresponding distribution

2 Experiment

2.1 Methodology

We have compiled a dataset of singing voices holding an f_0 of 300Hz for over 2 seconds. 300Hz was chosen as it was considered to be a comfortable frequency for both male and female subjects. The data was intended to represent a wide section of singing voices, this was necessary as our humanisation function has the intention of being non-specific in regards to variables such as genre, singing style, gender and singing experience. For this purpose, our subjects all held intentionally diverse levels of singing experience and training.

Recordings were taken using a MacBook Pro with a Mackie Onyx sound card and an AKG414 condenser microphone. Recordings were taken in a semi-anechoic chamber. Subjects were asked to listen to 300Hz sine wave at 50dB and simultaneously replicate the sound without vibrato. Subjects were able to monitor their own voice through headphones at roughly the same amplitude as the stimulus.

The recordings were then analyzed 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.

2.2 Observations

From our recordings, f_0 contours tend to have spaces in their distributions with clusters of sample points around smaller local mean values. Figure 1 demonstrates this concept with a 3 tier example of a tone being held. The vocalists f_0 changes sharply from one region to another, leaving three or four smaller clusters of sample points rather than a distribution that matches a global Gaussian curve.

All f_0 contours that were tracked had signs of this concept, the independent variables being the length and number of individual stable regions. Generally, those with more singing experience produced less regions, with a smaller amplitude or deviation from 300hz. For all vocalists, each of the regions have a varying local amplitude and no immediate trends are measured in the length of individual regions. The distribution of each region is considered to be Gaussian for synthesis purposes. By making this assumption, we are able to control the standard deviation (σ^2), mean (μ) and amplitude of each subsection, giving us additional control over parameters in the model. In future work, we aim to determine the validity of our assumed Gaussianity of individual regions in the model.

3 Modeling Drift

This work provides a system for mapping random normally distributed values to a statistically modeled distribution, based around the trajectory of the f_0 in human contours. The distributions are chosen empirically from observations made on our recorded dataset of amateur and semi-professional male and female vocalists producing an /a/ phoneme at specified frequencies.

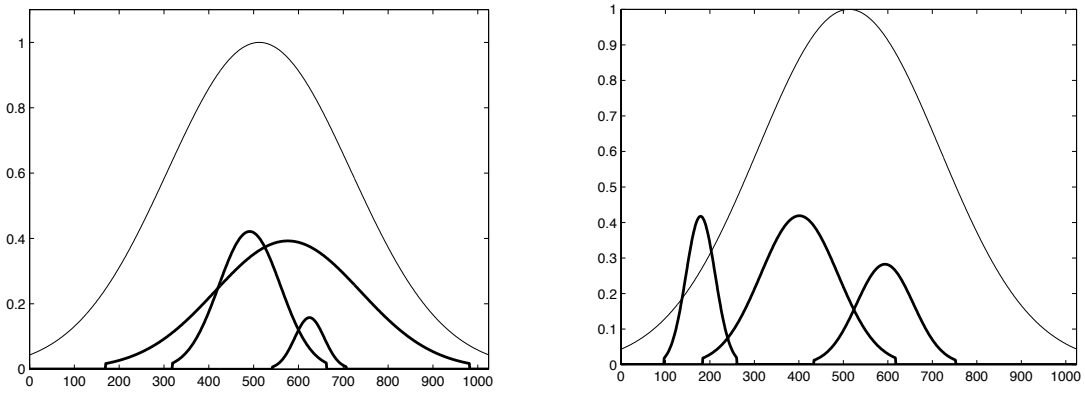

Figure 2: Examples of Distribution format

The suggested method is based around the drift that occurs as a result of auditory feedback. A proposed 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, it can be shown that rather than having a distribution, which fits that of white noise, smaller clusters of sample points at varying offsets from a mean frequency are observed. When represented by a Probability Density Function, we consider the signal to be a combination of smaller potentially, overlapping Gaussian curves, with regions of relatively low power in between.

The actual humanisation function is derived from the Box-Muller Transform [10], this quantizes a vector containing uniform values into a series of smaller, sliding and augmenting Gaussian curves with their corresponding standard deviations (σ^2) and offset mean (μ_n) values based on probability and human-error. The individual curves represent regions in the contour at which the f_0 is relatively stable, the mean frequency for each region (Xn) is assigned a random offset $_n$ value from the global mean, each corresponding standard deviation ($x\sigma^2$) is limited to 1/N of the global σ^2 , with N representing the number

of regions. The input to the function is low-passed filtered, uniformly distributed noise, with a cut-off frequency (f_c) of 5hz. After the values are generated, they are used as control signals for a PSOLA pitch shifting algorithm, in order to modify pitch quantized or synthetic singing voice.

The amount of smaller curves represent the distribution of the vocalists stable regions, which in turn is likely to represent the ability to hold a note. In future studies, we aim to estimate the correlation of these regions with the perceptual quality of the tested vocalists.

The function works by taking white noise signal which has been low pass filtered at 5Hz and shifting and scaling the distribution with weighted random values over a given time period. The pseudo-random process allows user- defined inputs such as the number of regions (or local mean values) and the amplitude of the pitch deviation. Each signal is then allocated a series of weighted pseudo-random points at which it is allowed to change its distribution.

$$\Delta f0 = x_t \sigma_i^2 \mu_i \quad (2)$$

Where: $\begin{cases} t \geq R_1, i = 1 \\ t \geq R_n, i = 2 \\ t \geq R_n, i = n \end{cases}$

Equation 2 shows the function used to create the humanized pseudo-random process where σ^2 ,μ_i and R_n represent vectors containing weighted random values for the standard deviation, mean and interval points respectively. x_t represents our filtered Gaussian input signal. Figure 2 provides examples of the system in operation. Each smaller Gaussian curve occurs at a different interval of t in time. Allowing the signal to drift between regions of relative stability. A comparison between the methods proposed in this paper can be seen in Figure 3.

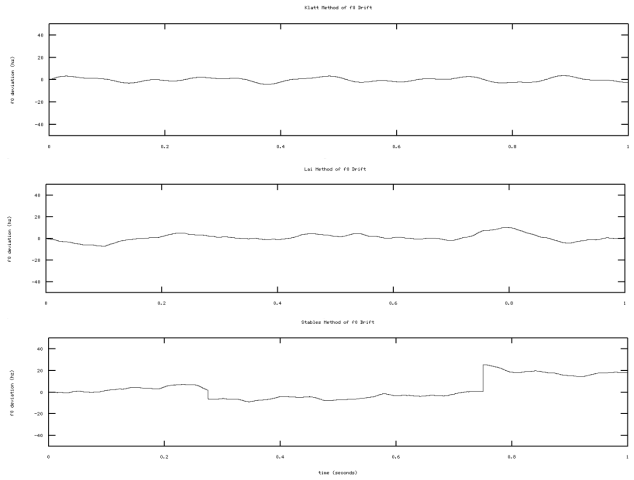

Figure 3: Comparison of techniques

4 Conclusion

In order to produce a drift signal, which mirrors that of the human f_0 control system, we apply a novel function to low pass filtered white 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 auditory feedback mechanism's influence over f_0 in singing, whilst 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.

From the research so far, we know very little regarding the lengths and distributions of the smaller regions. For the benefit of this synthesis experiment, we assume a level of Gaussianity. At this stage however, it is unlikely to be problematic due to the short lengths of time they are held for (around 1/3 second). From initial tests, we are able to produce singing voices which have a more natural, human like deviation in pitch when compared to existing systems. To quantify this however, listening tests are currently being implemented.

5 Future Work

We aim to make developments to the current model, for synthesis purposes each of our smaller regions of drift are produced using a Gaussian pseudorandom number generator. We aim to evaluate the validity of this model and to attach probability functions to the lengths, amplitudes and quantities of individual regions based on a wider dataset of singing voices. In future studies we aim to develop the humanisation function into a perceptual model by evaluating the parameters over multiple bands in a spectrum. We predict that varying levels of drift are required at different frequencies to give the perceived effect of a natural human singing voice. This is likely to correspond with the amount of smaller Gaussian regions and the global mean. Furthermore, we hope to augment our findings with auditory models in order to produce a new parametric approach to modeling pitch deviation in singing.

References

[1] R.F. Orlikoff and RJ Baken. Fundamental frequency modulation of the human voice by the heartbeat: preliminary results and possible mechanisms. The Journal of the Acoustical Society of America, 85:888, 1989.

[2] T.A. Burnett, J.E. Senner, and C.R. Larson. Voice F0 responses to pitch-shifted auditory feedback: a preliminary study. Journal of Voice, 11(2):202–211, 1997.

[3] A. Loscos and J. Bonada. Emulating rough and growl voice in spectral domain. In International Conference on Digital Audio Effects (DAFx04), Naples, Italy, 2004.

[4] L. Fabig and J. Janer. Transforming singing voice expression - the sweetness effect. 2004.

[5] M.W. Macon, L. Jensen-Link, J. Oliverio, M.A. Clements, and E.B. George. Concatenation-based MIDI-to- singing voice synthesis. PREPRINTS-AUDIO ENGINEERING SOCIETY, 1997.

[6] T. Saitou, M. Unoki, and M. Akagi. Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. Speech communication, 46(3-4):405–417, 2005.

[7] W.H. Lai. F0 Control Model for Mandarin Singing Voice Synthesis. IEEE, 2007.

[8] D.H. Klatt and L.C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. the Journal of the Acoustical Society of America, 87:820, 1990.

[9] X. Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on, volume 1. IEEE, 2005.

[10] ER Golder and JG Settle. The Box-Muller Method for generating pseudo-random normal deviates. Applied Statistics, 25(1):12–20, 1976.