THE VOWEL WORM: REAL-TIME MAPPING AND VISUALISATION OF SUNG VOWELS IN MUSIC

Harald Frostel, Andreas Arzt, Gerhard Widmer
Department of Computational Perception
Johannes Kepler University, Linz, Austria
harald.frostel@jku.at

ABSTRACT

This paper presents an approach to predicting vowel quality in vocal music performances, based on common acoustic features (mainly MFCCs). Rather than performing classification, we use linear regression to project spoken or sung vowels into a continuous articulatory space: the IPA Vowel Chart. We introduce a real-time on-line visualisation tool, the *Vowel Worm*, which builds upon the resulting models and displays the evolution of sung vowels over time in an intuitive manner. The concepts presented in this work can be used for artistic purposes and music teaching.

1. INTRODUCTION

An important aspect in singing is the production of distinct, recognisable vowels. The work presented in this paper aims to automatically recognise and track two perceptually important qualities ('open-/closeness' and 'front-/backness') of vowels in sung music, in real-time (and, by implication, recognising the vowels themselves). This would have many applications in science, music teaching, and art.

In the speech research and phonetics communities numerous studies have focused on the relationship between acoustic signal parameters of (spoken) vowels and their phonological categories and perceivable qualities (e.g., [1], [2], [3], [4], [5], [6], [7]). In particular, Pfitzinger [3], [4], [5], [6] has recently shown that such automatic mappings from acoustic to articulatory features are possible and can even match the performance of trained phoneticians.

With this paper, we wish to introduce the Sound and Music computing (SMC) community to that body of work and demonstrate that vowel qualities can be recognised also in vocal music performance. We first present systematic experiments that corroborate Pfitzinger's findings, on a different vowel corpus. In particular, we show that an effective mapping can also be learned on the basis of Mel Frequency Cepstral Coefficients (MFCCs) (which are routinely computed in many SMC applications). We then present an experimental tool that tracks and visualises sung vowels over time – as trajectories in a common phonological 'vowel space' (see Section 2) – in a real-time, on-line setting. In analogy to [8] we call this the *Vowel Worm*. (The figure in Section 5 and several demo videos (see Section 5) explain why.) It provides us with preliminary (though still somewhat anecdotal) evidence that this kind of mapping approach is indeed viable for the singing voice. [1]

The focus of the work presented here is not on categorical *classification* of vowels but on a mapping into a continuous (and phonologically motivated) two-dimensional space. Real-time categorical vowel recognition can easily be built on top of this – either via distance-based classification in the visualisation space or by modelling the vowel classes as MFCC distributions (e.g., in the form of Gaussian Mixture Models) and performing maximum-likelihood classification.

Our work is motivated by an artistic goal (real-time, on-stage music visualisation), but it could also be useful in music teaching – in particular, as a feedback tool in the training of singers, as briefly discussed in Section 6.

The paper is organised as follows. First, in Section 2, we explain the articulatory space we use for mapping and visualisation of vowel quality. In Section 3 we present our methods of modelling the projection into this space. Training and evaluation of these models are described in Section 4. We give insight into how this model was realised and used to visualise the vowel trajectory in Section 5. Finally, in Section 6, we discuss the results and possible scenarios where our approach might be applicable.

2. THE PERCEPTUAL SPACE OF VOWELS AND THE IPA VOWEL CHART

In the literature, mainly two kinds of diagram or chart are used to illustrate and classify articulatory vowel quality.

One type is the *formant frequency space* (e.g., [6], [2], [1]). The simplest version is a two-dimensional diagram with the first and the second formants $F1$ and $F2$ defining the axes, and vowels positioned in the diagram according to their formant frequencies. Variations exist with respect to the scaling of the formant frequencies (e.g., Hertz or Bark) and/or the usage of differences between formant frequencies instead of absolute values.

The second type of diagram is the *Cardinal Vowel Diagram* [11] and its newer adaptation by the International

[1] This is notable because in [9] and [6] it was shown that the fundamental frequency ($F0$) has a significant effect on the formant frequencies and, in particular, on the 'vowel height' (which is one of the articulatory dimensions we wish to recognise). In singing, this effect is expected to be much more pronounced than in speech, which was the focus of previous research.

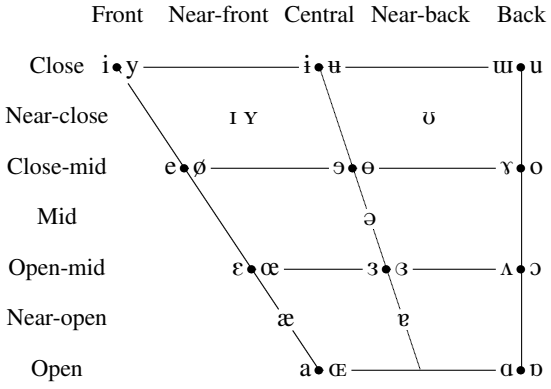

Figure 1. The IPA Vowel Chart [10]. The horizontal axis depicts the *vowel backness* and the vertical axis the *vowel height*. Vowels at the right and left of bullets are *rounded* and *unrounded*, respectively.

Phonetic Association (IPA), the vowel quadrilateral or *IPA Vowel Chart* [10] as shown in Figure 1. In our work we use the latter as a space to visualise vowel quality.

The IPA Vowel Chart locates vowels in terms of the tongue position required for their production. One dimension refers to the *backness* of the vowel, ranging from *front* to *back*. If the tongue or its highest point is placed near the front of the mouth (the hard palate), the vowel is labelled as a front vowel, whereas if the highest point of the tongue is placed at the back, narrowing the pharynx, the vowel is labelled as a back vowel.

The second dimension is *height*. If the tongue is near the roof of the mouth, the vowel height is described as *close*.[2] A vowel produced with maximum distance between tongue and palate is described as *open*. The eight primary cardinal vowels [i], [e], [ɛ], [a], [ɑ], [ɔ], [o], and [u] define the reference points in this chart. All other vowels can be placed in positions between them [10].

A third, partially independent dimension is the *lip rounding*. A vowel is called *rounded* if the lips are rounded during its production, and *unrounded* if the lips are relaxed.

3. VOWEL QUALITY PREDICTION

The objective of the work described here is to develop a system that can recognise, track, and visualise vowel qualities in sung music by mapping them into the IPA chart in real-time using a suitable regression model that is based on common audio features as routinely used in Music Information Retrieval (MIR) and SMC.

3.1 The Space

In order to build such a model, it is necessary to define a space on the basis of the IPA Vowel Chart. We simply place the vowel chart in a two-dimensional *Cartesian coordinate system* as shown in Figure 2. The proportions of the vowel chart used here are 2:3:4 for the bottom, right, and top sides respectively [12]. The backness coordinates of our space range from 0 (front) to 4 (back), and the height coordinates range from 0 (open) to 3 (close). Each vowel is therefore represented by a distinct point in this space.

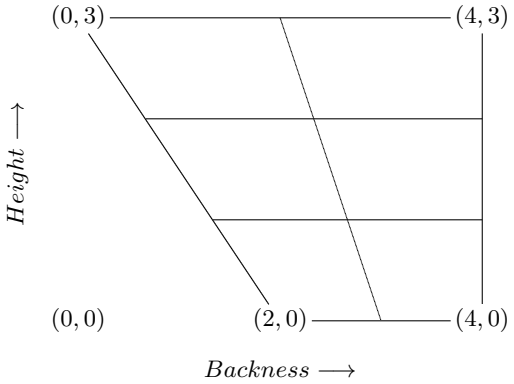

Figure 2. Coordinates of the vowel chart space used for regression with some sample points.

3.2 Multiple Linear Regression

As a predictor, we decided to use *multiple linear regression*. The advantages of this, compared to other methods like *local regression*, for instance, are its simplicity, robustness, and efficient implementation for real-time prediction.

$$y = \sum_i x_i \beta_i + \varepsilon = \langle \mathbf{x}, \boldsymbol{\beta} \rangle + \varepsilon \qquad (1)$$

As shown in equation 1, multiple linear regression is basically the inner product of a (feature) vector $\mathbf{x}$ (regressor) and a regression coefficients vector $\boldsymbol{\beta}$ plus an error term ε. It is necessary to extend $\mathbf{x}$ (the features used as regressors) by a further dimension to model the constant term in the linear equation. For each of the two vowel dimensions (backness and height) a separate regression model is generated.

3.3 Features

As regressors we use the following features (or subsets of these):

- **Mel Frequency Cepstral Coefficients (MFCCs) [13].** The MFCCs are calculated using 40 mel spaced triangular filters covering the spectrum up to 8000 Hz. No pre-emphasis is used, and the areas of the filters are not normalised; in other words, all filters have the same height. A Hamming window is used to calculate the spectrum.
- **Linear Prediction Filter Coefficients (LPFCs) [14].** Linear prediction of order 13 is performed, resulting in 14 filter coefficients.
- **Fundamental Frequency ($F0$).** To obtain the fundamental frequency (and to decide whether one is present at all), we use a (real-time) $F0$ estimation algorithm developed earlier and also implemented in our visualisation tool (Section 5). The underlying algorithm builds upon a combination of approaches

[2] Not close*d*, as one might expect from the naming of the other extreme ('open'). The vowel height dimension is sometimes also called 'closeness'.

presented in [15] and [16]. Several scalings of $F0$ are used, namely *Hertz*, *logarithm*, *mel scale*, and *equivalent rectangular bandwidth (ERB) scale*.

Altogether, this results in 58 features (40 MFCCs, 14 LPFCs, 4 representations of $F0$).

4. EXPERIMENTS

4.1 Corpus and Data Generation

We have not been able to find an annotated vowel database for sung vowels. Creating such a database is very labour-intensive, since in addition to recording the sung vowels, it also requires classification by several phoneticians who place them at the right positions in the IPA Vowel Chart, as done by Pfitzinger in [4] for spoken vowels. Thus, to build and validate our models, we relied on an existing database for spoken vowels.

We used the vowel corpus created for the *Vocal Joystick Project*[3] [17], which consists of a large amount of recorded monophthongs (pure vowel sounds with no changes in articulation) and vowel-to-vowel transitions (articulation moves from one vowel to another). The corpus features 9 vowels, namely [i], [e], [æ], [a], [ɑ], [o], [u], [ɨ], and the schwa [ə], spoken by multiple speakers (male and female with different native languages) and recorded at various sound levels, intonations, and durations [18]. In addition, each recording was judged by a phonetician as to whether it is of acceptable quality (i.e., close enough to the target vowel). Figure 3 shows the vowels covered by the corpus on the IPA Vowel Chart.

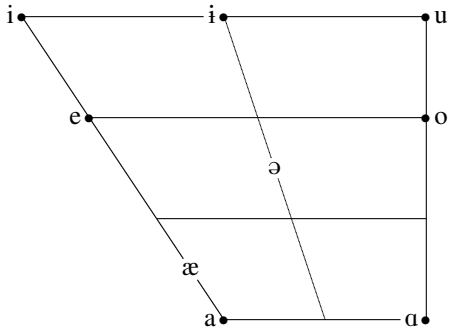

Figure 3. Vowels covered by the Vocal Joystick Corpus (cf. Figure 1).

For training and evaluation of our models, we used only utterances of monophthongs (no vowel-to-vowel transitions) of acceptable articulatory quality. After removing problematic utterances (i.e., those too far from the target vowel), we only kept speakers that still covered all 9 vowels. This resulted in a reduction from originally 92 speakers to 56. To create our data set, 150 uniformly distributed random points in time were generated for all corresponding utterances per speaker and vowel. Only voiced segments were taken into account, that is, segments with a detected fundamental frequency. This resulted in a total of 75,600 samples (56 speakers $\times$ 9 vowels $\times$ 150 samples). From these random points in time, the features (MFCCs, LPFCs, and $F0$) were extracted.

[3] Freely available at `http://www.vocaljoystick.org/`

4.2 Experimental Results

As mentioned above, our data covers 56 different speakers. We performed a 56-fold cross-validation for our evaluation. For every fold, the samples of one speaker were left out for testing while the rest of the data was used for training the regression models. After training, the models were used to predict the backness and height of the samples from this speaker, resulting in a *leave-one-speaker-out* cross-validation.

As evaluation measures we used the *correlation coefficient* r and the *root mean square error* (*RMSE*) for each of the two dimensions backness and height. The *RMSE* was calculated from the results in a normalised space, which means that the backness and height predictions were divided by 4 and 3, respectively, to obtain a chart space ranging from 0 to 1 in each dimension.

Features	t_{win}	r_b	$RMSE_b$
$\text{MFCC}_{1\text{-}40}$, LPFC	$93ms$	0.8854	16.09%
$\text{MFCC}_{1\text{-}40}$, LPFC, F0_{Hz}	$93ms$	0.8853	16.10%
$\text{MFCC}_{1\text{-}40}$, LPFC	$46ms$	0.8826	16.27%
$\text{MFCC}_{2\text{-}25}$	$93ms$	0.8659	17.32%
$\text{MFCC}_{2\text{-}25}$, F0_{ERB}	$93ms$	0.8656	17.33%
$\text{MFCC}_{2\text{-}25}$	$46ms$	0.8608	17.61%
$\text{MFCC}_{2\text{-}13}$	$46ms$	0.8572	17.82%
$\text{MFCC}_{1\text{-}6}$	$23ms$	0.8130	20.15%
Baseline	–	$1.6e$-13	34.61%

Table 1. Features and results for backness regression models sorted by r_b.

Features	t_{win}	r_h	$RMSE_h$
$\text{MFCC}_{1\text{-}40}$, LPFC, F0_{Hz}, F0_{Log}, F0_{ERB}, F0_{Mel}	$93ms$	0.8554	20.36%
$\text{MFCC}_{1\text{-}40}$, LPFC, F0_{ERB}	$93ms$	0.8551	20.38%
$\text{MFCC}_{1\text{-}40}$, LPFC	$93ms$	0.8540	20.45%
$\text{MFCC}_{2\text{-}25}$, F0_{ERB}	$93ms$	0.8526	20.53%
$\text{MFCC}_{2\text{-}25}$, F0_{ERB}	$46ms$	0.8502	20.68%
$\text{MFCC}_{2\text{-}40}$	$93ms$	0.8501	20.69%
$\text{MFCC}_{2\text{-}25}$	$46ms$	0.8479	20.83%
LPFC, F0_{ERB}	$93ms$	0.6701	29.17%
Baseline	–	$-8e$-17	39.28%

Table 2. Features and results for height regression models sorted by r_h.

We tried several window sizes, zero-padding sizes, and feature combinations to determine which settings perform best. Tables 1 and 2 show the results for several feature combinations and window sizes (t_{win}) but list only a subset of all tested combinations. Zero-padding does not seem to have any significant influence on the quality of the MFCCs. For performance reasons, we thus omitted zero-padding in the final results. The window size, however, does have an

impact: the larger the window, the better the regression. This is not surprising, as only monophthongs were used for training, and therefore a larger window covers more information and might cancel out variations in the signal. Furthermore, if a sample happens to be at the beginning or end of a vowel, a larger window captures more of the relevant signal. The baselines shown in Tables 1 and 2 represent a predictor that outputs the average backness and height.

The best *backness* correlation ($r_b = 0.8854$) was obtained with a window size of 93 ms, all MFCCs and LPFCs as features. Adding the fundamental frequency did not lead to any improvement. Without LPFCs, the best results were obtained with MFCCs 2 to 25 (excluding the first coefficient). The prediction of backness seems to be very robust in terms of the feature combinations. The worst result with only 6 MFCCs and a window of 23 ms still achieved a correlation coefficient of $r_b = 0.8130$.

Vowel *height* seems to be more sensitive to the choise of feature combination. Again, the best result was obtained with all features, but also including all scalings of $F0$. The height predictions improved with the usage of $F0$. This finding is in agreement with Pfitzinger's results [3], [5], [6], which showed that $F0$ influences the perceived vowel height. However, the impact is rather small in our case. Also, the scaling of $F0$ has no major influence. The best results were obtained by using the $F0$ in ERB scale. The reason for the limited effect of the fundamental frequency might be that the MFCCs themselves encode some amount of pitch information.

For comparison, in [5], Pfitzinger achieved results of $r_b = 0.964$, $r_h = 0.903$ and $r_b = 0.965$, $r_h = 0.960$ without and with $F0$, respectively. However, it is difficult to compare these results with ours. Pfitzinger used only 12 German speakers and different vowels. The vowel stimuli were presented to 40 trained phoneticians, each of whom assigned the stimuli to precise positions in the vowel chart (whereas we were limited to assigning training vowels to their grid positions in the chart, because we only had vowel labels as ground truth). It could thus be argued that the training data used in [4] is substantially more refined than the data available to us.

Note that vowel quality assessment is generally not unambiguous. Dioubina and Pfitzinger [12] stated that the judgement of vowel quality by phoneticians is influenced by their native language. In addition, according to [4], even skilled phoneticians cannot reliably repeat their own judgements after some time.

Overall, we conclude from the experiments for the purposes of our project that (1) the qualities backness and height of spoken vowels can be predicted reasonably well by linear regression, and (2) this can be done by using MFCCs, which are routinely computed and used in MIR and SMC applications as underlying audio features. Whether these results can be extended to *sung* vowels in music will need to be established quantitatively in future experiments – should an annotated corpus of sung vowels becomes available. In Section 5, we give some anecdotal evidence of this by supplying examples of our vowel visualiser in action.

4.3 Final Model

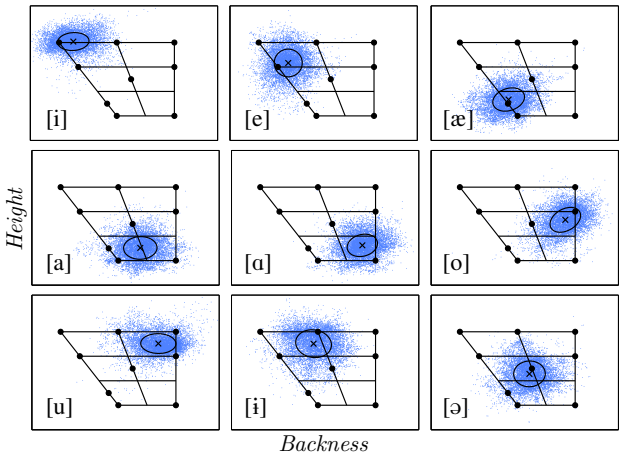

Figure 4. Scatter plots of the predictions of the final model for all 9 vowels (cf. Figure 3). Each dot marks the predicted point of one test sample. Each cross and circle indicates the mean and standard deviation of all predictions of one class.

We used 24 MFCCs (coefficients 2 to 25) with a window size of t_{win} = 46 ms as the final model parameters in the backness prediction. For height, two models were generated: one based only on the same 24 MFCCs as for backness, and one based on the 24 MFCCs plus the $F0$ in ERB scale. The system switches automatically between the models, depending on the presence of a valid $F0$. The chosen models constitute a trade-off between run-time and accuracy. A longer signal window raises computational costs in the calculation of the spectrum and the MFCCs, and also generally the tracker's latency. An effect similar to that of a larger window can be obtained by smoothing the final predictions of backness and height over time. Figure 4 shows scatter plots of the predictions of the trained models.

5. THE VOWEL WORM

Our ultimate goal is the visual tracking of vowels (and other aspects of singing) in artistic musical contexts. From previous experiments with categorical vowel classification, we learned that a simple textual display of the currently recognised vowel is not very useful. Vowel changes happen too fast for the viewer to process and validate these impressions. Also, classifying vowels into nominal classes has several drawbacks. Spoken or sung vowels are almost always a mixture of adjacent vowel classes. Moreover, classification limits the set of vowels to those used in training. If there are, for instance, only 9 vowel classes in the training data (as in the Vocal Joystick corpus), the classifier will recognise only those. With regression, on the other hand, only a subset of vowels is needed to develop a model that is (at least in theory) capable of projecting any vowel or input into a vowel space, even if it was not available during training.

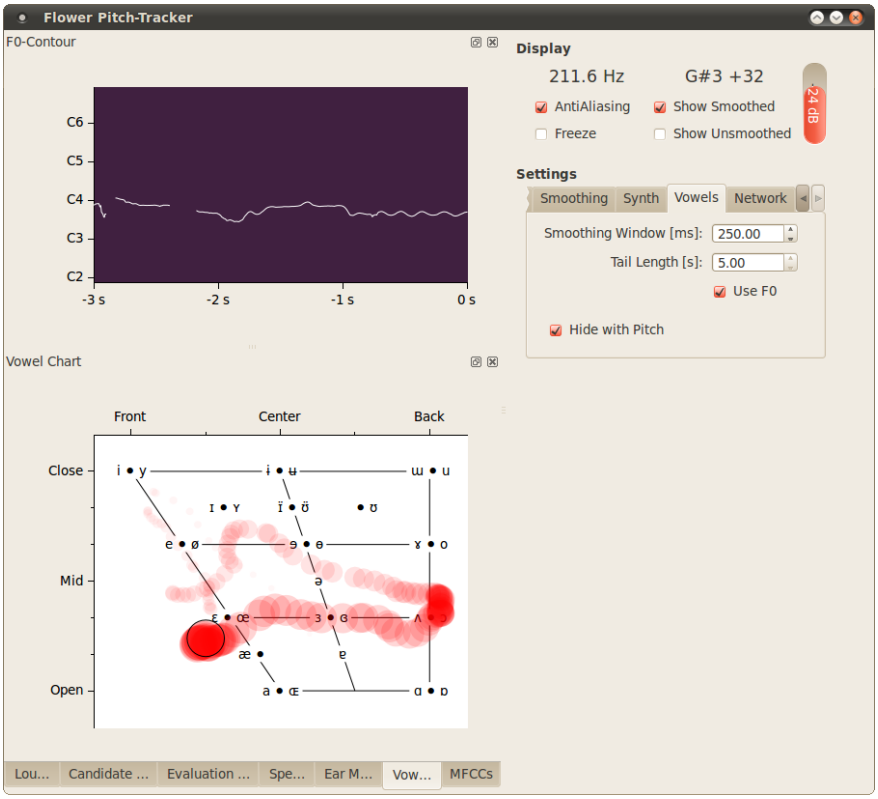

Figure 5. Screenshot of the real-time visualisation of the Vowel Worm. The upper left panel depicts the pitch contour of the last 3 seconds – as computed by our on-line pitch tracker. The botton left panel shows the trajectory of the sung vowels (Vowel Worm). On the right, the Vowel Worm settings can be adjusted.

These considerations led us to developing the *Vowel Worm*.[4] It not only displays the current vowel in a more intuitive manner by positioning it in the vowel chart plane, it also captures the evolution of the sung vowels over time. The Vowel Worm uses the regression models described above to perform real-time mapping of an incoming audio signal (ideally a solo singing voice) onto the IPA Vowel Chart visualisation plane. The current position in the chart is indicated by a filled circle, while instances further in the past appear smaller and fainter. Figure 5 shows a screenshot of our visualisation tool during a live performance. The top left panel shows the estimated $F0$-contour for the last 3 seconds, as determined by our pitch tracker. The Vowel Worm is located at the bottom left.

The Worm takes the MFCCs and the fundamental frequency $F0$ (computed on-line) as input. It then calculates backness and height via the model's regression formula. Depending on the presence of a valid $F0$, the implementation chooses the appropriate regression model for predicting the height. Backness and height are further smoothed over time. Smoothing is done for each dimension separately by averaging over the last predictions. This improves the visual stability of the projection, which otherwise tends to jitter. Typical smoothing windows range from 150 ms to 350 ms. Other settings that can be adjusted during runtime are the tail length of the worm (i.e., the time span into the past that is visualised), whether $F0$ is to be used, and whether the worm should hide in the absence of a fundamental frequency.

To give the reader an impression of the visualisation, we have generated a few *screen shot videos* of the Worm in action. They can be found at `http://www.cp.jku.at/projects/realtime/vowelworm.html`. In these examples, the input audio stream comes from an audio file, but the system works in exactly the same way with real-time input via microphone, for instance.

6. CONCLUSIONS AND DISCUSSION

This paper has addressed the problem of real-time vowel quality recognition and tracking, and introduced a particular way of mapping and visualising the development of sung vowels over time. The main result, as we see it, is that two central articulatory features of (spoken) vowels seem to be reliably predictable from standard MFCC features and that – on the basis of our unsystematic qualitative experiences with the Vowel Worm – models learned from spoken vowels appear also to apply (perhaps unexpectedly) well to sung vowels. Carrying out systematic quantitative experiments to support this would require the availability of precisely annotated musical corpora.

The concrete goal of the present project is to develop real-time music analysis and tracking technology that can be used to control live on-stage visualisations of large musical works (e.g., operas). To that end, we are also working on

[4] Concept and name were inspired by previous work on real-time visualisation of expressive performance parameters in the *Performance Worm* [8].

tracking other parameters (including exact score position) in, for instance operatic singing. For artistic visualisation purposes, the current recognition capabilities of the Vowel Worm may be considered sufficient.

Beyond artistic visualisation projects, we envision application in a vocal quality visualiser, particularly in didactic settings, and as a feedback tool for the training of singers, actors, etc. A prerequisite for this, however, would be precise quantitative experiments that establish whether, and to what extent, the placement predicted by learned models are reliably correct, which in turn depends on the availability of high-quality training and validation corpora of sung vowels (of various musical styles).

Acknowledgments

This research is supported by the City of Linz, the Federal State of Upper Austria, and the Austrian Research Fund (FWF) under grant TRP109-N23. The FLOWER real-time audio processing framework, upon which the Vowel Worm is built, is being developed by Martin Gasser (Austrian Research Institute for Artificial Intelligence, Vienna), supported by FWF project Z159.

7. REFERENCES

[1] W. Klein, R. Plomp, and L. C. W. Pols, "Vowel Spectra, Vowel Spaces, and Vowel Identification," *The Journal of the Acoustical Society of America*, vol. 48, no. 4B, pp. 999–1009, 1970.

[2] M. Aylett, "Using Statistics to Model the Vowel Space," in *Proceedings of the Edinburgh Linguistics Department Conference*, 1996, pp. 7–17.

[3] H. R. Pfitzinger, "Dynamic Vowel Quality: A new Determination Formalism based on Perceptual Experiments," in *4th European Conference on Speech Communication and Technology (EUROSPEECH '95)*, vol. 1, Madrid, Spain, Sep. 1995, pp. 417–420.

[4] ——, "Acoustic Correlates of the IPA Vowel Diagram," in *Proceedings of the 15th International Congress of Phonetic Sciences*, vol. 2, Barcelona, Spain, Aug. 2003, pp. 1441–1444.

[5] ——, "The /i/-/a/-/u/-ness of Spoken Vowels," in *8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 809–812.

[6] ——, *Speech Production and Perception: Experimental Analyses and Models*. Berlin: ZAS Papers in Linguistics, 2005, vol. 40, ch. Towards Functional Modelling of Relationships between the Acoustics and Perception of Vowels, pp. 133–144.

[7] S. Ran, B. Millar, and P. Rose, "Automatic Vowel Quality Description using a Variable Mapping to an Eight Cardinal Vowel Reference Set," in *Proceedings of the 4th International Conference on Spoken Language (ICSLP 96)*, vol. 1, Oct. 1996, pp. 102–105.

[8] S. Dixon, W. Goebl, and G. Widmer, "The Performance Worm: Real Time Visualisation of Expression based on Langner's Tempo-Loudness Animation," in *Proceedings of the International Computer Music Conference (ICMC)*, Göteborg, Sweden, Sep. 2002, pp. 361–364.

[9] J. Sundberg, *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[10] International Phonetic Association, *Handbook of the International Phonetic Association : A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Jun. 1999.

[11] D. Jones, *An Outline of English Phonetics*, 9th ed. W. Heffer & Sons Ltd., 1962.

[12] O. I. Dioubina and H. R. Pfitzinger, "An IPA Vowel Diagram Approach to Analysing L1 Effects on Vowel Production and Perception," in *7th International Conference on Spoken Language Processing*, vol. 4, Denver, Colorado, USA, Sep. 2002, pp. 2265–2268.

[13] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[14] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals," in *Proceedings of the 1967 IEEE Conference on Communication Processing*, 1967, pp. 360–361.

[15] A. Camacho and J. Harris, "A Pitch Estimation Algorithm Based on the Smooth Harmonic Average Peak-to-Valley Envelope," in *IEEE International Symposium on Circuits and Systems (ISCAS 2007)*, May 2007, pp. 3940–3943.

[16] H. Frostel, "Real-time Fundamental Frequency Estimation of the Human Voice," Master's thesis, Johannes Kepler University, Linz, Austria, 2009.

[17] J. A. Bilmes, X. Li, J. Malkin, K. Kilanski, R. Wright, K. Kirchhoff, A. Subramanya, S. Harada, J. A. Landay, P. Dowden, and H. Chizeck, "The Vocal Joystick: A Voice-Based Human-Computer Interface for Individuals with Motor Impairments," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 995–1002.

[18] K. Kilanski, J. Malkin, X. Li, R. Wright, and J. Bilmes, "The Vocal Joystick Data Collection Effort and Vowel Corpus," in *Proceedings of the International Conference on Spoken Language Processing*, Pittsburg, Pennsylvania, USA, Sep. 2006.