Audio Engineering Society

Convention Paper

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

Feature extraction for voice-driven synthesis

Jordi Janer[1]

[1]*IUA-MTG, Universitat Pompeu Fabra, Barcelona*

Correspondence should be addressed to Jordi Janer (jjaner@iua.upf.es)

ABSTRACT

This paper explores the singing voice from an unusual perspective, not as a musical instrument but as a musical controller. A set of spectral processing algorithms extract features form the input voice. These features are categorized in four groups: excitation, vocal tract, voice quality and context. The extracted values are then transmitted as Open Sound Control (OSC) messages in order to be used in an external synthesis engine. In this document, we provide first a technical description of the algorithms, and in a second part, we detail the components of the system. A practical example of voice-driven synthesis using PureData (Pd) is also presented.

1. INTRO

The singing voice as musical instrument has played a prominent role in the history of music since ancient times. Its main characteristics are probably the high expressiveness and its ubiquity. In this paper, the aim is to use the voice to control other sounds, to capture vocal gestures. Essentially, vocal gestures can be of three different nature in the voice production mechanism. They can have its origin in the breathing system (airflow), glottis (vocal folds), and the vocal tract (timbre). Regarding the analysis of the incoming voice signal the source/filter model describes, to some extent, what is going on in the voice production mechanism, e.g. pitch, vocal folds tension and formant/articulators setup. Based on that, the use of spectral techniques can help us in extracting attributes from the voice.

Certainly, in the past decade audio content description has become a relevant topic in the field of Music Technology. Mainly, the goal of these approaches is to analyze and to add some metadata to an audio material. This metadata can be used a posteriori integrated in music retrieving systems such as music recommenders.

On the other side, Digital Musical Instruments have widen the possibilities in terms of sound production and the control over this production. These systems are no more bound up to a certain physical construction. Concerning the control, the MIDI protocol has become the standard either in commercial or experimental systems.

In the approach presented in this paper, we address both topics: sound description and musical control transmission protocols. The solution adopted uses Open Sound Control (OSC), and escapes from the limitation of the MIDI messages.

2. FEATURE EXTRACTION

A preliminary issue that arises is the selection of attributes that will be estimated, in other words, how we parameterize the voice. Moreover, the chosen attributes or features should be controllable by the user and, to some extent, intuitive. In this approach, we propose a set of features categorized in four groups: Excitation, Vocal Tract, Voice Quality and Context. The feature extraction work flow consists of first, a signal windowing and the Discrete Fourier Transform. In a further step, an harmonic analysis of the spectrum, based on Phase-Locked Vocoder [3] is achieved. With this additional information, the algorithms estimate various voice attributes.

2.1. Excitation Descriptors

This *excitation* category refers to the train of pulses generated in the glottis, which depends on the airflow pressure and the vocal folds tension. We can link these two with output signal energy and fundamental frequency, respectively.

The energy estimation procedure takes a block of audio samples and calculates the mean square. Estimating the pitch from the voice might be a difficult task, since the pitch may present rapid and constant changes. On top of that, during the note attack (note onset), the signal may have an instantaneous pitch far away from the final stabilized pitch. Early techniques were developed for speech processing [5], but we employ here a method primarily focused on the singing voice. Our implementation, includes a Pitch Estimation method based on the Two-Way-Mismatch technique [1].

2.2. Vocal Tract descriptors

It seems also interesting for our work to extract timbre information from the input voice. Later, these attributes can modify characteristics of the synthesized sound. In the human voice, the timbre is related to the pronounced vowel, which is determined by the formants. Hence, it will be very intuitive for the user to execute a *musical gesture* by changing from one vowel to another. We provide here a very basic method for estimating the formant frequencies that are necessary for identifying a vowel.

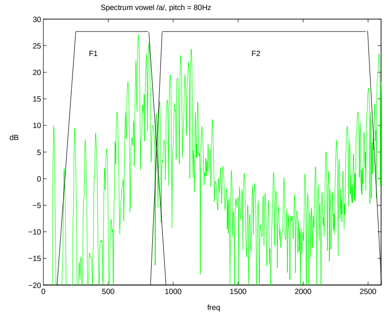

Fig. 1: Spectrum of the vowel /a/ with a pitch of 80Hz. Superimposed the weighting functions for finding $F1$ and $F2$.

This algorithm separates the spectrum into two regions, one for the first formant and one for the second formant, and is based on the spectral centroid calculation.

As reported in [4], the vowels are spread over a triangle in a two-axis representation of first and second formant. From this representation, we can make some assumptions. Typical values for $F1$ are between 200 and 800Hz, and $F2$ goes from 800 to 2500Hz. We calculate the spectral centroid within each region, which will be directly our estimated values for $F1$ and $F2$. However, we use a slightly modified equation for calculating $F1$ and $F2$. Firstly, instead of taking all FFT bins, we consider only spectral peaks magnitude ($A[k]$) and frequency ($f[k]$); and secondly, the magnitude $A[k]$ is weighted by a“crossover-like” function $w_i[k]$ for splitting smoothly the two regions. Finally, we end up with two equations,

$$F1 = \frac{\sum_{k=1}^{N} A[k] w_1[k] f[k]}{\sum_{k=1}^{N} A[k] w_1[k]} \tag{1}$$

$$F2 = \frac{\sum_{k=1}^{N} A[k] w_2[k] f[k]}{\sum_{k=1}^{N} A[k] w_2[k]} \tag{2}$$

In fact, this method works pretty good for most vowel sounds, when the formants frequencies are separated and fall in the right region. However, we find also sounds in which the two formants are very close, and both in the first region. Particularly, this is the case of the vowel /u/ and sometimes a *close* /o/. A possible solution would be to compute the energy ratio between the two regions ($E_{Region1}/E_{Region2}$), to detect these cases. And then, to apply a second split in the first frequency region and calculate the centroids of the subregions. Note also that for high pitched voices, the formants are undersampled by the harmonics peaks, invalidating the estimatied values.

2.3. Voice Quality descriptors

The human phonation process may suffer of artifacts that alterate voice quality. In a medical context, these artifacts are considered as voice disorders. In contrast, singers apply often these artifacts as effects (roughness, growl, etc.) that add expression to the performance. In this sense, we aim to detect the presence of these artifacts in a sustained voiced sound and use them as control parameters. We developed two algorithms, *Sub-Harmonic Factor (SHF)* and *Partial Frequency Stability (PFS)*, that describe, to some extent, roughness and breathy effects.

We addressed first the *Sub-Harmonic Factor* algorithm. In a normal situation, the vocal folds vibrate at a certain vibrating frequency. It means that a constant cycle of opening and closing is periodically repeated. However, if this cycle does not remain constant, either in its amplitude or shape, new frequency components appear in the signal, producing hoarse voice. These new components are in fact sub-harmonics of the fundamental vibrating frequency. The *Sub- Harmonic Factor* algorithm is described in detail in [2], and the result tends to 1.0 for a high sub- harmonicity, and is 0 in case of a signal with solely pure harmonics.

On the other side, the *breathy* voice is caused by a particular mode of phonation. The physiological explanation of breathiness can be found in [6]. According to Sundberg, there is a *phonatory dimension* ranging from pressed to breathy phonation. When a high sub-glottal pressure is combined with a high degree of adduction, a pressed phonation occurs. If the sub-glottal pressure decreases, jointly with a lower adduction, it is called a*flow phonation*. Then, if the adduction force is still reduced, the vocal folds fail to make contact, producing the known breathy phonation.

Perceptually, a breathy phonation results into an harmonic spectrum mixed with high-frequency noise, due to the air turbulence. Hence, our goal is to examine the harmonic content at high frequencies, and to extract a valid *breathiness* descriptors. Taking also the output of the SPP Analysis, this method observes the trajectories of the detected harmonic partials within a frequency range from 3 to 6kHz. Due to the presence of noise, the harmonic peaks' frequencies suffer of great deviations in the mentioned range. For a frame k, we calculate a frequency stability factor (S_f), which are the difference between the measured frequency ($f_r[k]$), and the predicted one ($\hat{f}_r[k] = n \cdot pitch$).

$$S_{fr}[k] = |f_r[k] - \hat{f}_r[k]| \tag{3}$$

Finally, we correct the summation with the delta-pitch factor $d_{pitch}[k]$, being k the frame index (equation 5). The factor $d_{pitch}[k]$ will ensure that the final value is not affected by normal pitch variations such as in a vibrato.

$$d_{pitch}[k] = 1 - 100\,\frac{pitch[k] - pitch[k-1]}{pitch[k] + pitch[k-1]} \tag{4}$$

$$PFS[k] = d_{pitch}[k]\frac{1}{R}\sum_{r=0}^{R} S_{\phi,r}[k] \tag{5}$$

2.4. Context descriptors

Up to this point, we have addressed the analysis of the voice by extracting a set of instantaneous parameters frame by frame. However, expressive gestures might not be instantaneous but being extended over a time period. Therefore it seems convenient to make observation over the time, and try to identify possible gestures.

As an example of a context descriptors, we propose the *Attack Unvoiceness* descriptor. Here we seek to capture the characteristics of the note's attack. In our context, we assume that a *note* consists of a voiced sound, a vowel, with a rather constant pitch. However, by preceding the note with a consonant (unpitched sound) preceding the sustained voiced sound, we can make use of additional information, and map it in a further step to the synthesized sound. Hence, our model of note, will have two parts: an unpitched attack plus the actual pitched note. Basically, this descriptor aims at extracting the harshness of the note attack. For each frame k, it takes a block of samples, and computes the *Energy Factor*, and the *Zero-Crossing Rate*. An instantaneous attack factor $A[k]$ is calculated using the formula 6, and it is added to an accumulator variable $A_{context}[k]$.

$$A[k] = \{ \begin{array}{ll} 0 & , pitched \\ \sqrt{Energy[k] \cdot ZCross[k]} & , unpitched \end{array} \tag{6}$$

3. IMPLEMENTATION

An indispensable requirement of such a system is to work in real- time, so that it can be used as musical instrument. Spectral processing techniques, which require a large computational capacity, can run nowadays real-time on standard desktop computers. We implemented a system consisting of a C++ stand-alone application, which basically receives an audio stream, analyzes the signal, and sends UDP (User Datagram Protocol) packets. The voice is captured by a microphone and is transferred to the application using ASIO drivers. The voice features are extracted and transmitted at a particular frame rate. In this implementation we use a window size of 2048 samples, a FFT size of 4096 and a hop size of 256, giving a frame rate of 172 fps, which is acceptable for controlling sound synthesis parameters with enough resolution.

In the field of digital musical instruments, alternatives to the MIDI protocol have emerged over the last years. One of the most relevant is Open Sound Control, developed at CNMAT[1]. Basically, the OSC protocol allows to send UDP packets to a client, in this case, the synthesis engine.

Without going into mapping considerations, which are beyond the scope of this paper, we present a practical example of using the system jointly with a synthesis software. A simple FM generator, implemented in PureData, receives OSC messages from the voice analysis application. Here, the voice features control frequency and amplitude of the oscillators. The corresponding PureData patch is shown in figure 2 The observed latency is around the 30 ms, which is sufficient for synthesizing continuous varying sounds.

4. CONCLUSIONS

The objective of this work was to develop algorithms that capture expressive aspects of the singing voice beyond the *energy* and *pitch*. A set of analysis algorithms that extract features from the singing voice in spectral domain is introduced. In a second step, these features are transformed into OSC messages, which allows the integration in various music software systems. A basic example is presented, whereby singing voice attributes such as energy, pitch and formant frequencies drive the parameters of a simple FM synthesis engine. Although the estimation of some features needs to be more robust, we consider that this approach results in a rewarding musical experience.

This research has been partially funded by the EU-FP6-IST- 507913 project SemanticHIFI.

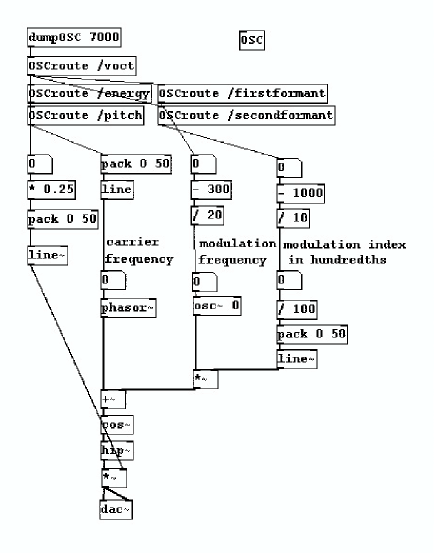

Fig. 2: This patch shows an example of using extracted voice attributes for controlling a simple FM synthesis in PureData.

5. REFERENCES

[1] P. Cano. Fundamental frequency estimation in the SMS analysis. In *Proceedings of COST G6 Conference on Digital Audio Effects 1998,* Barcelona, 1998.

[2] L. Fabig and J. Janer. Transforming singing voice expression - the sweetness effect. In *Proceedings of the 7th. International Conference on Digital Audio Effects (DAFX),* Naples, Italy, 2004.

[3] J. Laroche and M. Dolson. New Phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 1999.

[4] D. O'Shaughnessy. *Speech Communication, Human and Machine.* Addison-Wesley, New York, 1987.

[5] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

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

[1] http://www.opensoundcontrol.org/cnmat/