A Pilot Study for Algorithmic Diction Detection for Use by Singers and Vocal Teachers

Bhawna Rathi, Timothy Y. Hsu

Department of Music and Arts Technology, IUPUI
brathi@iu.edu, hsut@iu.edu

Abstract. This paper introduces an algorithmic signal processing method to quantify vocal diction using audio files that can potentially assist singers and teachers. Clear diction and pronunciation in singing is important for a variety of reasons and should be exercised alongside the development of voice. In order to convey a clear verbal message, strong diction is needed. To accomplish this goal of diction detection, the interpretation of the consonants is of prime significance. The proposed algorithm works with features such as zero crossing rate, spectral spread, spectral flux and spectral centroid. In this paper, we offer a proposed framework and algorithm of diction detection using modern applicable audio features and extraction techniques. Future approach for analysis of diction is also defined.

Keywords. Diction, Signal processing, Vocal technologies, Detection, Audio feature extraction

1 Introduction

Although there are several systems for pitch and speech recognition using signal processing, few works have been conducted to measure diction. The accuracy of the sung text is an important aspect of the enjoyment of vocal music by audiences and a core concern for singers. Strong diction of the text in vocal music provides the textual narrative. If the words cannot be understood, the song can become meaningless [1], [2]. The diverse singing genres of classical/opera, music theater, and popular music necessitates that singers produce a wide variety of tonal qualities. For classical singers, they must find an equilibrium between vocal resonance and diction over a wide range of pitches. For music theater performers, the drama of the text demands that the singer communicates the script while sometimes sacrificing tonal quality for better diction. In the popular music genre, the desire for sound can be organized widely, from gritty and chaotic to mellow, calm and romantic, resulting in varying diction needs. In this set of musical demands, the diction ranges from seemingly unintelligible to crystal clear. Because of these wide articulatory demands, there is a need for a quantitative method of determining vocal diction to potentially track the impact of vocal coaching/lessons in modern voice studios. The purpose of this work is to include a framework and algorithm that can detect and assess the quality of vocal

diction so that students and vocal coaches can potentially receive informative feedback [3].

2 Background

Active research has been ongoing for decades to distinguish, recognize and spot vowels and stop consonants, as well as to detect speech. For example, speech segmentation, speech-verification, prosody modification and emotion conversion use the onset and offset point detection to investigate the excitation of sources, the spectral peaks, and the modulation spectrum for the determination of the vowel offset point, as well as the spectral energy of glottal closure region [4],[5]. Several algorithms have been shown to correctly detect the vast majority of vocal onsets in fluent speech. Most algorithms are based on the assumption that vocal onsets are marked by the appearance of rapidly growing resonance peaks. Whereas there has been active research with artificial intelligence on algorithms for speech recognition (i.e. Google Voice, Apple Siri, Amazon Alexa), specific research for detecting vocal diction that focuses on the strength of consonants rather than detecting the vowels and speech is lacking [6].

Consonants, as opposed to vowels, requires abrupt changes in the articulators and constriction of the vocal tract. These abrupt changes occur through stops, aspiration, and friction. The articulation of consonants are affectded by voicing, place of articulation, and manner of articulation. Additionally, the production of consonants allow for consonants to be categorized into stops, approximants, nasals, and affricates. Stops require a full closing of the vocal tract by the lips, tongue, or glottis. In stops, there is a quiet interval accompanied by a sudden sound onset. Depending on where the flow is stopped, different voice and unvoiced consonants are produced, such as /b/, /p/, /d/, /t/, /k/, /g/. Fricatives, or continuants, are produced by pushing air through a small constriction along the vocal tract. This constriction causes turbulence, creating noise as the air passes through the constriction. The constriction can occur at the teeth or teeth, resulting in voice and unvoiced sounds such as /s/, /z/, /f/, /v/. In approximants, the articulators involved in the constriction are further apart than in the case of fricatives, but still alter the flow as compared to a neutral state. This constriction, however, does not produce turbulence, so these consonants do not feature noise. Examples of approximants are /w/, /j/, /l/. Nasals, like stops, entirely restrict the air flow through the mouth and instead pass the air through the nasal cavity. The tongue or lips can be used to stop the air flow, resulting in /n/ and /m/ consonants. With affricates, there is a quiet interval accompanied by more prolonged noisy onset. These affricates share characteristics of both stops and fricatives where there is a full air blockage, a release, and then a noise-like fricative. Examples of affricates are the *ch* and *j* sound [7].

As vocal music is set based on poetry, storytelling, or narration, it is important that the lyrics be distinguished through good diction. Some vocalists, however, may unintentionally sing with less consonant precision, leading to poorer textual communication to the audience. Poor singing diction may result from improper mouth

shape while singing, inappropriate tongue placement, and poor breath control. Improving diction is a can improve the communicative musical experience, where a consistent diction can deliver text resulting in the voice becoming a persuasive instrument [1]-[3].

Signal Processing Background

Acoustic analysis and digital signal processing has been utilized in speech analysis, music information retrieval, nondestructive testing, and other fields. For music information retrieval methods, it is common to employ preprocessing, time framing, windowing, and short time Fourier transforms to calculate features and onset detections [8]-[12]. Much of the previous research in singing analysis has focused on pitch detection and formant analysis, other features are needed to analyze consonants and they are described here. The spectral flux (SF) estimates the amount of change in the spectral shape. It is defined as the average difference between short time Fourier Transform (STFT) frames. Low SF can suggest events such as a steady state feedback of signal, a pause in the signal, or the pitch shift associated with a new note. The spectral centroid represents the center of gravity of the spectral energy. It is the frequency-weighted sum of the power spectrum normalized by its unweighted sum. The spectral spread represents the intensity of the power spectrum along the spectral centroid and describes the spectral shape. It can be defined as a standard deviation of the power spectrum along the spectral centroid. The zero-crossing rate (ZCR) is defined as the number of sign changes in sequential blocks of audio samples. To evaluate ZCR, each pair of consecutive samples has a positive of negative sign that determines the zero-crossing probability of the signal [8].

3 Methodology

MATLAB, with the Audio Toolbox, is the primary computational package used as it contains a variety of signal processing and statistical tools. As shown in Fig. 1, after loading the audio file, the signal is split into overlapping frames using a Hamming window and filtered to remove low frequency. For this paper, a 1000ms Hamming window, with 60 percent overlap, was used. The audio features, as described above, are then extracted for each frame. The purpose of the feature extraction is to obtain a compact representation of these salient acoustic characteristics of the signal. The amplitudes of each of the feature's sets are normalized by dividing them by their maximum value. The individual extracted features are merged into a combined resultant feature by performing a weighted sum of all the calculated features, where the weighting value allows for a prioritization of some features more than others. This weighted sum results in the "Calculated Diction." The weighting factors were determined experimentally to produce values that correspond with diction detection. Additionally, to focus solely on consonants, vowel suppression is needed. By suppressing the portions of sustained singing vowels, individual timbral differences between singers is mitigated, thus focusing the algorithm on consonants only. To separate the vowels and consonants, the gradient of spectral centroid was calculated. When the gradient is zero (or near-zero), it is assumed that these are moments that

vowels are being sung. Thus, in these time periods where the gradient is near zero, the Calculated Diction is suppressed, resulting in a consonant focused method. Additionally, because some consonants are short time events (/t/ and /k/) and others are longer (/s/ and /m/), short time averages are calculated to magnify wide and/or elided consonants. Finally, the results are plotted to visually represent the quality of diction over time.

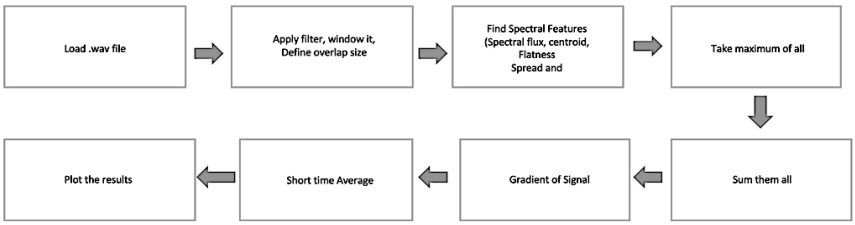

Fig. 1. Flow diagram of diction algorithmic process.

4 Results

The experimental diction analysis was conducted on several audio files in order to determine effectiveness and robustness of this proposed framework. For the Calculated Diction, the final weighting factors were that the spectral flux was multiplied by 0.5, spectral centroid was multiplied by 1, spectral spread was multiplied by 0, and the zero-crossing rate was multiplied by 0.5. Two recordings of a male vocalist singing: “My country, ’tis of thee, Sweet land of liberty, of thee I Sing” were used. The vocalist sang the text twice, with the first iteration utilizing “good” diction and the second iteration using “bad” diction. In Fig. 2, the Calculated Diction for the “good” recording shows sharp peaks for /c/, /tr/, /s/, /l/, and /t/, whereas the Calculated Diction for the “bad” recording shows fewer and weaker overall peaks. For example, the /tr/ sound of “Country” has a clearer peak for the “good” recording as compared to the “bad” recording. Future work will implement filters that mitigate noise arise due to the breathing noise that may be problematic in longer phrases that require breathing. A second audio file of “Mary had a little lamb” was analyzed in the same manner as the above, with both “good” and “bad” recordings. In Fig. 3, it can be seen in the Calculated Diction, that /m/, /h/, /d/, and /t/ have prominent peaks in the “good” recording, but the /r/ sound is entirely missing in the Calculated Diction for the “bad” recording. While upon initial inspection, it may seem like the /l/ is more present in the “bad” recording, this could potentially be explained by the normalization process of the Calculated Diction as normalization currently occurs within each individual sound file, rather than through a common reference. It can also be noted that there are false positive peaks in the “bad” side for both Calculated Diction and the short time average plots. The presence of these false positive may actually be indicative of poor diction, where vowels begin to blend with consonants. Because of this blending, the threshold used to suppress vowels becomes more unclear; thus, in the future, adaptive thresholding may improve the results.

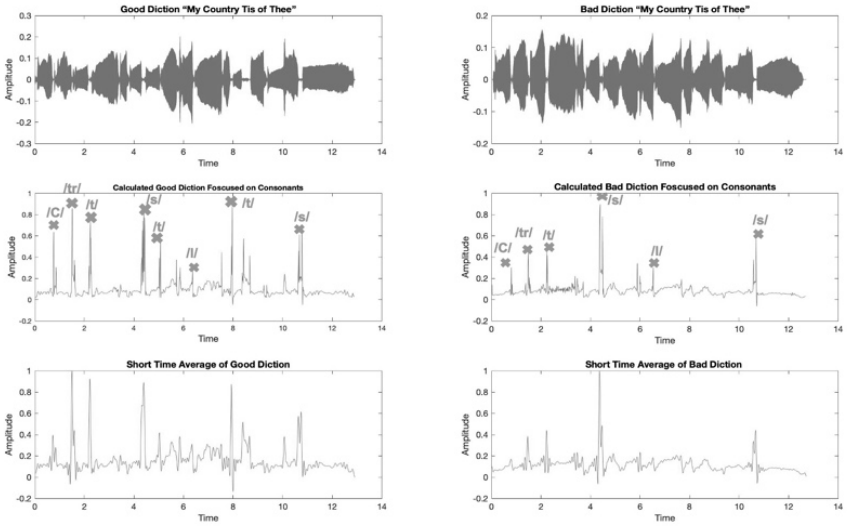

Fig. 2. Graphical results of "My country, 'tis of thee, Sweet land of liberty, of thee I sing".

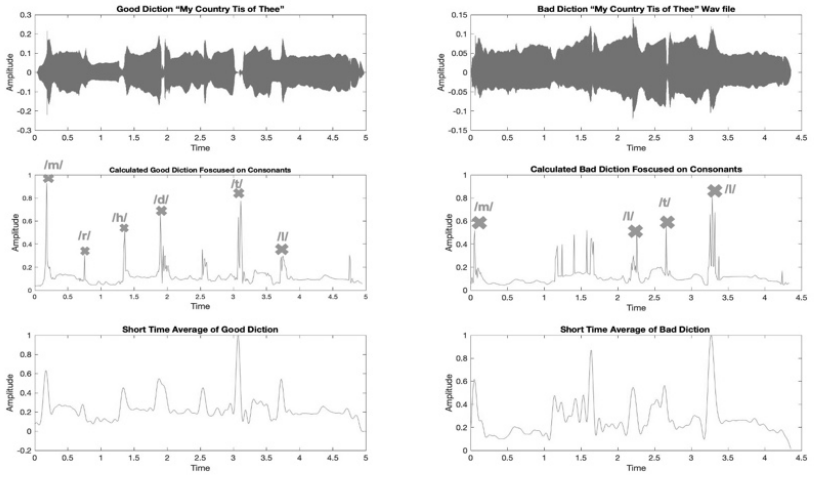

Fig. 3. Graphical results of "Mary had a little lamb".

5 Discussion

This proposed algorithm results in a Calculated Diction that shows promising results in being able to visually differentiate between good and bad diction. In particular, appropriate peaks for stop consonants are present, as stops are normally preceded by a

silent sudden interval which is followed by an abrupt onset of sound. The results suggest that this method may not be as sensitive for other consonants like /l/ and /m/, potentially due to the voiced and non-percussive nature of the production. Additionally, through the weighting, it is seen that the zero-crossing rate and spectral flux seem to do better at detecting the consonants, whereas the spectral spread and spectral flatness do not contribute as much to the weighted sum. In practice, students and vocal coaches can use these figures as visual aids to show not only overall diction, but also to reveal precise time stamps where the Calculated Diction does not result in an expected peak. Future research will include improvement in the accuracy through formant analysis and machine learning, a real-time implementation, and a potentially user-friendly "Diction Meter" that can be adopted by teachers and students.

6 Conclusion

The algorithm is still under development as accuracy of consonant detection and thresholds can continue to improve. This framework is unique in that it prioritizes diction, and intentionally suppresses typical singing features such as pitch, vowels, and individual timbres. The ultimate goal of this project is to create a user-friendly interface for singers and vocal coaches to receive real-time visual feedback on the quality of diction, leading to better communication of text and a more refined overall musical experience.

References

[1] A. David, "A Handbook of Diction for Singers: Italian, German, French Catalog", *New York : Oxford University Press, 20, pp 44-51, 2008.*

[2] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler, "A tutorial on onset detection in music signals," in *IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035-1047, 2005.*

[3] J. Ginsborg & F.Philip & B.Christopher, "Have we made ourselves clear? Singers and non-singers' perceptions of the intelligibility of sung text" *International Symposium on Performance Science, 2005.*

[4] J. Yadav and K. S. Rao, "Detection of Vowel Offset Point From Speech Signal," in *IEEE Signal Processing , vol. 20, no. 4, pp. 299-302,2013.*

[5] R. Jeremiah, L. M. (n.d.)., "Predicting vowel and consonant confusions using signal processing techniques,," in International Congress Series, *vol.*1273, *pp. 15-18, 2004.*

[6] F. Toa, J. H. L Hansen, C. Busso "An unsupervised visual-only voice activity detection approach using temporal orofacial features", *In INTERSPEECH, 2015.*

[7] E. D. Casserly, P.David "Speech perception and production", *Wiley Interdisciplinary Reviews: Cognitive Science. vol.1, pp.* 629 - 647*, 2010.*

[8] M. Banitalebi-Dehkordi, A. Banitalebi-Dehkordi , "Music Genre Classification Using Spectral Analysis and Sparse Representation of the Signals.," in *Journal of Signal Processing Systems vol 74, 2014.*

[9] S. E. Blumstein, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," in *The Journal of the Acoustical Society of America* vol *66, no. 1001,1979*.

[10] H. L. Tan, Y. Zhu, L. Chaisorn and S. Rahardja, "Audio onset detection using energy-based and pitch-based processing," Proceedings of *2010 IEEE International Symposium on Circuits and Systems, Paris, France, pp. 3689-369, 2010.*

[11] M.Bhattacharjee S.Prasanna, S., & P.Guha, Time-Frequency Audio Features for Speech-Music Classification. ArXiv, abs/1811.01222, 2018.