S-Transform Features for Automatic Classification of Phonation Types

A Doondi Bala Saraswathi*, DVNSL Praharshitha*, Kandi Aditri Shreeya*, Purva Sharma*, Gurugubelli Krishna†
* Department of Electronics and Communication Engineering
Koneru Lakshmaiah Education Foundation, Hyderabad, India
2320040050@klh.edu.in, 23200040059@klh.edu.in, 2320040015@klh.edu.in, dr.purvasharma@klh.edu.in
† Samsung Research and Development Institute, Bengaluru, India
krishna.g@samsung.com

***Abstract*—Automatic classification of phonation type is an important task in speech and singing voice analysis. Different phonation modes, such as breathy, neutral, and pressed, provide essential information about the movement of the vocal folds and the characteristics of voice production. In order to capture variations in glottal cycle along with formant variation for different phonation types, this study explores the Stockwell Transform based cepstral features (STCC) for the classification of different phonation types. The performance of the proposed feature was also compared with the baseline features such as Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Zero-Time Windowing Cepstral Coefficients (ZTWCC), Wavelet Transform Cepstral Coefficients, and Single Frequency Cepstral Coefficient (SFFCC). Compared to baseline features, the proposed features outperformed them in terms of classification accuracy, F1-score, and AUC. The results of the experiment show that the proposed STCC features improved the accuracy of phonation classification by approximately 5% and 3% on the Finnish vowels and the soprano singing datasets compared to the baseline features.**

***Index Terms*—S-Transform, Cepstral coefficients, Phonation types, Singing voice analysis, Support vector machine.**

I. Introduction

Human voice production is governed by the interaction between the excitation generated by the vocal folds and the filtering effect of the vocal tract system. Variations in vocal fold vibration, airflow, and laryngeal tension produce different voice qualities [1], commonly referred to as phonation modes. These phonation modes form a continuum of expressive variations and play an important role in singing, speech communication, emotional expression, and clinical voice evaluation. In singing voice production [2], phonation characteristics strongly influence vocal timbre, expressiveness, and perceived vocal effort, making phonation analysis an important task in speech processing, vocal pedagogy, and medical voice evaluation. Therefore, automatic categorization of phonation modes has become an important research problem in speech and singing voice analysis.

Previous studies have identified several phonation modes such as breathy, neutral, flow, and pressed, each characterized by different degrees of vocal fold adduction, airflow leakage, and muscular tension [3]. Breathy phonation is typically associated with incomplete glottal closure and turbulent airflow, whereas pressed phonation involves strong adduction of the vocal folds and increased laryngeal tension [4]. Neutral and flow phonations lie between these two extremes and often exhibit similar acoustic characteristics, making their automatic classification more challenging.

Accurate identification of phonation modes is required in several applications, including singing voice synthesis, expressive speech generation, speaker analysis, and clinical evaluation of voice quality [5]. In [6], authors explored low frequency spectral density (LFSD) as feature for classifcation of breathy, modal, and pressed phonation types. Maxima dispersion quotient (MDQ) was shown to be robust feature for discrimination of breathy voice from tense voice in presence of 10 dB SNR in [7]. Various glottal parameters such as normalized amplitude quotient (NAQ), closing quotient, quasi-open quotient (QOQ), open quotient and the speed quotient were explored in [8] for analyzing the different voice qualities. Recent studies explored different time-frequency methods such as wavelet-scattering network (WSN)-based features [9], zero-time windowing (ZTW) [10], and single frequency filtering (SFF) [11], [12] method based cepstral features for classification of phonation types in singing voice.

Despite the promising performance that these methods have already shown, the ability to capture glottal excitation and formant transition remains a problematic issue, especially when it comes to singing voice phonation analysis. From previous work on classification of voice disorder [13], it was observed that S-Transform based [14] time-frequency representation captures variation in the glottal cycle along with formant transition more effectively than other time-frequency representations. With this motivation, this study proposes S-Transform Cepstral Coefficients (STCC) for automatic classification of different phonation types. Performance of the proposed features was also compared with baseline features such as MFCC [15], [16], ZTWCC [15], SFFCC [17] and WTCC [9] on Finnish vowel [8] and the Soprano singing dataset [18]. The experiments were performed using a Support Vector Machine classifier using a quadratic kernel.

The rest of the paper is organized as follows. Section II describes phonation analysis through Stockwell Tranform, Section III presents experimental setup with details of fea-

tures extraction, database, and classifier. Results of automatic detection of phonation types are presented in IV. Finally, conclusions of this work are discussed in Section V.

II. Phonation Analysis Through S-Transform

In order to reliably categorize different phonation modes [1], it is necessary to analyze the acoustic signal using an appropriate time–frequency representation that accounts for both excitation and vocal-tract characteristics. The nature of phonation in both speech and singing is determined by the vibratory behavior of the vocal folds together with the interactive coupling between the vocal folds and the vocal tract [4]. Variations in vocal-fold adduction rate, airflow, and laryngeal tension give rise to different phonation modes, which manifest as non-stationary spectral patterns in the acoustic signal‘ [13]. Consequently, phonation analysis requires time–frequency representations that are capable of capturing both short-duration excitation events and long-term spectral context. The S-Transform [14] has been found to be effective for such analysis due to its adaptive time-frequency resolution.

A. S-Transform Representation

Phonation signals contain temporal and spectral variations; hence, the S-Transform is used in this study to obtain an appropriate time–frequency representation. The S-Transform [14] is a time-frequency analysis method that combines the phase-preserving property of the Short-Time Fourier Transform with the multiresolution capability of wavelet analysis. Considering a continuous-time signal, $x(t)$, the S-Transform, denoted as $S_x(\tau, f)$, is defined as

$$S_x(\tau, f) = \int_{-\infty}^{\infty} x(t)\, g(\tau - t, f)\, e^{-j2\pi f t}\, dt \tag{1}$$

where $x(t)$ represents the signal, f is the normalized frequency, and S denotes the power spectrum of the signal at that frequency, characterized by the total near-end signal energy of all its constituent components. Here, τ denotes time, f denotes frequency, and $g(\tau - t, f)$ is a frequency-dependent Gaussian window. The Gaussian window is defined as

$$g(t, f) = \frac{1}{\sqrt{2\pi}\,\sigma(f)} \exp\left(-\frac{t^2}{2\sigma^2(f)}\right) \tag{2}$$

where the standard deviation $\sigma(f)$ is inversely proportional to frequency and is given by

$$\sigma(f) = \frac{1}{|f|} \tag{3}$$

This formulation allows the analysis window to automatically vary with frequency. At lower frequencies, a wider window improves frequency resolution, whereas at higher frequencies, a narrower window enhances temporal resolution [19]. As a result, the S-Transform is able to simultaneously capture slowly varying spectral characteristics, such as formant structures, as well as rapidly varying temporal events associated with vocal-fold excitation [13].

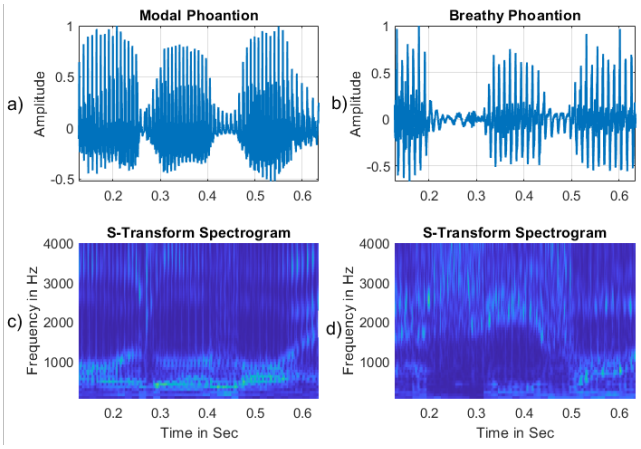

Fig. 1. Illustration of speech signal and spectrograms obtained with S-Transform for modal and breathy phonation respectively. (a)-(b) Speech signal. (c)-(d) Stockwell-Transform based spectrogram.

Fig. 1 illustrates the speech signal for modal and breathy phonation and its corresponding S-Transform based spectrogram. From the figure, we can observe that formant transition as well as energy variation in the glottal cycle effectively. For breathy phonation the S-transform also captures the high frequency harmonics, which arises due to air leakage through the vibrating vocal folds. The examples shown in Fig. 1 are included to provide a qualitative illustration of the time-frequency characteristics captured by the S-Transform for different phonation modes. Although the selected examples may differ in their fundamental frequency, the objective of the figure is to highlight the differences in spectro-temporal structure rather than pitch. Furthermore, the classification results reported in this work are obtained from complete datasets containing multiple speakers and a broad range of pitch values, thereby reducing the influence of speaker- or pitch-specific variations on the overall findings.

B. Cepstral Feature Extraction of S-Transform

The S-Transform representation is then used to derive cepstral coefficients so that a compact set of features can be obtained for phonation analysis [13]. Considering a discrete time speech signal $y[n]$, the speech signal can be pre-emphasised to boost the high-frequency components, and the pre-emphasised signal is taken as $x[n]$ as determined by

$$x[n] = y[n] - 0.9y[n-1] \tag{4}$$

The pre-emphasised signal is further divided into short time frames, with a duration of 20 ms, and the S-Transform is computed for each 20 ms segment to obtain the time–frequency representation of the signal [19]. In order to extract cepstral features of the S-Transform, the magnitude spectrum is initially calculated and logarithmic compression is applied to reduce the dynamic range. Spectral component decorrelation is then performed using the discrete cosine transform (DCT), and the cepstral coefficients are computed as

$$c(k,n) = DCT\left\{\log|S_x(n\Delta T, m\Delta f)|\right\} \quad (5)$$

where $S_x(n\Delta T, m\Delta f)$ denotes the S-Transform of the speech signal.

Lastly, the initial 13 static cepstral coefficients along with their respective delta and delta-delta coefficients are computed, yielding a 39-dimensional feature representation known as S-Transform Cepstral Coefficients (STCC), which is used for phonation classification.

III. Experimental Setup

This section presents the experimental setup used to evaluate the proposed features along with the baseline feature sets. Details of the feature extraction process, database, and classifier used for automatic classification of different phonation modes in both speech and singing voice signals are described in this section.

A. Features

In this study, MFCC [15], LPCC [16], ZTWCC [10], SFFCC [17], and WTCC [13] are used as baseline features along with the proposed STCC features for automatic detection of phonation types.

- MFCC and LPCC features are computed using a frame length of 20 ms and a hop length of 5 ms. For each frame, 13 static, 13 delta, and 13 delta-delta coefficients are extracted, resulting in a 39-dimensional feature vector. For these 39 coefficients, statistical measures such as mean, standard deviation, skewness, and kurtosis (stastical averages) are computed, which results in a 156-dimensional feature vector.
- In order to capture spectro-temporal variations of speech signal Single frequency filter (SFF) method [11], decomposes the speech signal into multiple frequency bands. ZTW method focuses on fine spectral details by using impulse-like windowing to obtain instantaneous spectral information with high temporal resolution [20]. Wavelet-transform [7] provides good frequency resolution at low frequency and good time resolution at high frequency. Speech signal is decomposed into its time-frequency componnets using SFF, ZTW and WT methods. From these components, 39 dimension (13 static, 13 delta, and 13 delta-delta) cepstral coefficients are computed and referred to as SFFCC, ZTWCC and WTCC, respectively.
- The proposed STCC features are obtained by computing the S-Transform of the speech signal as described in Section II-B.

B. Database

Two datasets are used in this study, namely the Finnish vowel dataset and the Soprano singing dataset .

- Finnish vowel dataset contains eight vowels of breathy, pressed and modal phonation. Database was recorded with sampling frequency of 44.1 kHz. It contains a total of 792 recordings by 11 speakers [8].
- Soprano database is the singing vowel dataset [18], It was recorded by a professional Russian soprano female singer in modal, breathy, pressed and flow phonation types. Total of 763 voice samples with sampling frequency of 96 kHz are available with scale being change from A3 to G5.

All recordings from the Finnish vowel and Soprano singing datasets were resampled to 8 kHz prior to feature extraction. This was done to maintain a common sampling rate across both datasets and to reduce computational requirements during feature computation. Since the phonation-related information considered in this study is primarily reflected in the lower-frequency region of the spectrum, the selected sampling rate is sufficient to preserve the acoustic characteristics required for phonation analysis while ensuring consistency across all experiments.

C. Classifier

Automatic classification of phontaion types was carried out using different machine learning classifiers such as logistic regression, k-nearest neighbour, decision tree, support vector machine, medium neural network and narrow neural network. Among all the evaluated classifiers, the SVM with a quadratic kernel provided the best overall performance. Therefore, it was selected for all subsequent experiments. The classifier is trained to distinguish between the phonation categories present in the datasets, namely breathy, neutral, and pressed in the Finnish dataset, and breathy, neutral, flow, and pressed in the Soprano singing dataset . To ensure an unbiased evaluation, 10-fold cross-validation is adopted, and the reported results correspond to the average performance across all folds. The evaluation is carried out using accuracy, recall, and F1-score to reflect overall classification performance and the balance between detection and prediction. In addition, the Area Under the ROC Curve (AUC) is considered to assess the separability of phonation classes, with values closer to 1 indicating improved discriminative capability. The classification experiments were carried out at the utterance level. Cepstral features were first extracted from short-time frames of each recording. To obtain a fixed-dimensional representation suitable for classification, statistical measures including mean, standard deviation, skewness, and kurtosis were computed over the frame-level features. The resulting utterance-level feature vectors were subsequently used for training and evaluation of the classifier.

IV. Results

TABLE I and TABLE II depicts the results obtained for the automatic classification of phonation types on Finnish vowel dataset. The classification results are reported in terms of accuracy, recall, F1-score, and AUC. Performance is reported for baseline features such as MFCC, LPCC, ZTWCC, SFFCC and WTCC along with proposed STCC features.

Among the baseline methods, SFFCC achieved the highest classification accuracy of 73.8%. This indicates that the SFF spectrum is capable of capturing phonation-related characteristics effectively. It can also be observed from the TABLE I,

STCC features gave the best performance in terms of accuracy, recall, F1-score and AUC of 77.4%, 77.4%, 77.4% and 0.96, respectively, for Finnish vowel database. This result suggests that the S-Transform representation contains discriminative information that is useful for distinguishing among different phonation types.

TABLE I
PERFORMANCE COMPARISON USING BASELINE FEATURES AND STCC FEATURES IN TERMS OF ACCURACY, RECALL, AREA UNDER THE ROC CURVE (AUC), AND F1-SCORE ON FINNISH PHONATION DATASET

Feature	Accuracy (%)	Recall (%)	F1 (%)	AUC
MFCC	73.1	73.07	73.08	0.95
LPCC	74.6	74.60	74.60	0.92
ZTWCC	62.3	62.33	62.31	0.89
SFFCC	73.8	73.83	73.81	0.94
WTCC	69.62	69.62	69.65	0.94
STCC	**77.4**	**77.4**	**77.4**	**0.96**

TABLE II shows the performance analysis on Soprano singing dataset . The proposed STCC feature set provides the best performance, achieving an accuracy of 82.2%, recall of 80.9%, F1-score of 80.9%, and AUC of 0.98. The better performance suggests that the S-Transform representation preserves variations related to vocal-fold excitation as well as formant structure more clearly. These characteristics play an important role in differentiating phonation modes, which results in higher classification accuracy.

TABLE II
PERFORMANCE COMPARISON USING BASELINE FEATURES AND STCC FEATURES IN TERMS OF ACCURACY, RECALL, AREA UNDER THE ROC CURVE (AUC), AND F1-SCORE ON SOPRANO SINGING DATABASE

Feature	Accuracy (%)	Recall (%)	F1 (%)	AUC
MFCC	78.8	76.68	77.73	0.97
LPCC	70.2	68.90	69.54	0.90
ZTWCC	64.4	62.13	63.24	0.92
SFFCC	79.8	77.57	78.67	0.98
WTCC	73.1	74.03	73.57	0.93
STCC	**82.2**	**80.9**	**80.9**	**0.98**

The confusion matrices presented in Fig. 2 provide additional insight into the behaviour of the proposed STCC features across individual phonation categories. For the Finnish vowel dataset, breathy phonation achieved the highest recognition performance among the three classes. This observation suggests that the acoustic effects associated with airflow leakage and breathiness are represented effectively within the extracted feature space. Most of the classification errors occur between neutral and pressed phonation, which is consistent with the gradual transition that exists between neighbouring phonation categories.

A similar pattern can be observed for the Soprano singing dataset. Pressed phonation achieved the highest class-wise

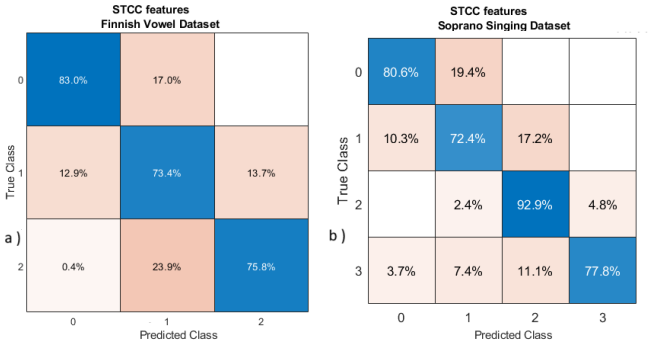

Fig. 2. Illustration of Confusion matrix (a)-(b) Confusion matrix for STCC features on Finnish Vowel and Soprano singing dataset , respectively. The horizontal and vertical axis shows the predicted classes and true classes, respectively. Class labels 0, 1, 2, and 3 represent breathy, neutral, and pressed and flow phonation type, respectively.

recognition performance, indicating that the excitation characteristics associated with stronger vocal-fold constriction are captured reliably by the proposed representation. The remaining confusions are primarily observed between acoustically similar phonation types, particularly those located adjacent to one another along the phonation continuum. Despite these overlaps, the overall results demonstrate that STCC features provide sufficient discriminative information for reliable phonation classification in both speech and singing voice signals.

To better understand the observed improvement in classification performance, the characteristics captured by the STCC representation are further discussed in the following subsection.

A. Interpretation of STCC Features

The performance improvement observed with STCC features appears to be closely related to the way the S-Transform represents the underlying speech signal. Differences among phonation types arise from changes in vocal-fold vibration, glottal closure behaviour, airflow leakage, and laryngeal tension. These physiological variations influence both the excitation source and the spectral characteristics of the produced speech signal, giving rise to distinct acoustic patterns.

In breathy phonation, incomplete glottal closure allows a greater amount of air to pass through the glottis, introducing additional turbulent components into the signal. Pressed phonation, in contrast, is associated with stronger vocal-fold adduction and more abrupt glottal closure, often resulting in a more pronounced harmonic structure. Neutral phonation lies between these two extremes and therefore shares acoustic characteristics with both categories, which can make its automatic classification more challenging.

A key advantage of the S-Transform is its ability to adapt its analysis window according to frequency. Lower-frequency regions are analysed with improved frequency resolution, whereas higher-frequency regions benefit from improved temporal resolution. This property allows the representation to

capture both slowly varying vocal-tract information and rapid excitation-related events within a unified framework.

The results obtained in this study suggest that preserving both source-related and vocal-tract-related information is beneficial for phonation analysis. Compared with several conventional cepstral representations, STCC features appear to retain phonation-dependent characteristics more effectively, leading to improved discrimination among breathy, neutral, pressed, and flow phonation types across both speech and singing voice datasets.

V. Conclusion

This paper presents an STCC-based representation for the analysis and classification of phonation modes in speech and singing vowel signals. The performance of the proposed features is compared with baseline features such as MFCC, LPCC, ZTWCC, SFFCC, and WTCC. All experiments were carried out on publicly available Finnish vowel and soprano singing vowel datasets using an SVM classifier.

From the experimental results, it is observed that STCC features provide the best performance for automatic classification of phonation types in both speech and singing vowel data. The highest classification accuracy of 78% and 81.7% was obtained on the Finnish and Soprano singing dataset s, respectively. The improved performance may be attributed to the ability of the S-Transform to capture phonation-related characteristics more effectively than other time–frequency based methods.

Different phonation modes arise due to the interaction between the oscillation of the vocal folds and the resonant filtering of the vocal tract. The S-Transform is capable of capturing variations in the glottal cycle along with formant variations, which are important for discriminating between different phonation types. Overall, the results indicate that an effective time–frequency representation plays a key role in extracting phonation-related information from speech signals.

References

[1] J. Laver, *The phonetic description of voice quality*. Cambdrige University Press, 1980.

[2] J. Sundberg, "The perception of singing," in *The psychology of music*. Elsevier, 1999, pp. 171–214.

[3] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[4] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Communication*, vol. 118, pp. 33–47, 2020.

[5] J.-L. Rouas and L. Ioannidis, "Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings," in *interspeech*, vol. 2016, 2016, pp. 150–154.

[6] D. N. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density." in *INTERSPEECH*, 2013, pp. 3206–3210.

[7] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1170–1179, 2013.

[8] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types." in *INTERSPEECH*, 2007, pp. 1410–1413.

[9] K. R. Mittapalle and P. Alku, "Tunable q-factor wavelet transform-based features in the classification of phonation types in the singing and speaking voice," *Journal of Voice*, 2024.

[10] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs)." in *Interspeech*, 2018, pp. 232–236.

[11] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[12] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6410–6414.

[13] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Stockwell-transform based feature representation for detection and assessment of voice disorders," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 101–119, 2024.

[14] R. G. Stockwell, L. Mansinha, and R. P. Lowe, "Localization of the complex spectrum: the s transform," *IEEE transactions on signal processing*, vol. 44, no. 4, pp. 998–1001, 2002.

[15] S. R. Kadiri and P. Alku, "Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. EL418–EL423, 2019.

[16] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. Gómez-Vilda, "Lpc, lpcc and mfcc parameterisation applied to the detection of voice impairments," *tc*, vol. 2, p. 1, 2000.

[17] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (sffcc)." in *Interspeech*, 2018, pp. 441–445.

[18] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[19] H. K. Vydana and A. K. Vuppala, "Detection of fricatives using s-transform," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3896–3907, 2016.

[20] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.