CONSTANT Q CEPSTRAL COEFFICIENTS FOR CLASSIFICATION OF NORMAL *VS.* PATHOLOGICAL INFANT CRY

Hemant A. Patil, Ankur T. Patil, Aastha Kachhi

Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology (DAIICT),
Gandhinagar, India.

ABSTRACT

Classification of normal *vs.* pathological infant cry is an interesting and technologically challenging research problem due to *quasi-periodic* sampling of vocal tract spectrum by high pitch-source harmonics resulting in extremely poor spectral resolution for commonly used spectral features, such as Mel Frequency Cepstral Coefficients (MFCC). To that effect, in this paper, we propose a new approach of feature extraction based on Constant Q Transform (CQT) that is known to have variable spectro-temporal resolution w.r.t Heisenberg's uncertainty principle in signal processing framework. Further, CQT is also known to preserve *form-invariance* property (than its Short-Time Fourier Transform (STFT) counterpart)- a desirable attribute of feature descriptors to be invariant w.r.t shape, shift, rotation, and scaling. CQT- based features are then transformed to the cepstral-domain to derive Constant Q Cepstral Coefficients (CQCC), which are then fed to statistical and discriminative classifiers, namely, Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) respectively. CQCC-GMM and CQCC-SVM systems gave relatively better results than MFCC for various experimental evaluation factors for infant cry classification task on widely used and statistically meaningful Baby Chilanto Database. Relatively best performance, in particular, 99.82% accuracy (0.44% EER), is observed for CQCC-GMM system.

Index Terms— Infant Cry Classification, Constant Q Transform, Short-Time Fourier Transform, GMM, SVM.

1. INTRODUCTION

Infant cry analysis and classification is highly interdisciplinary in nature involving physiological, neurological, pediatrics, engineering, developmental linguist, and psychology [1]. Around three million infants die within first month after the birth, which may be due to vaccine preventable diseases, other pathologies, and malnutrition. In this context, recently fingerprint-based biometrics is developed for infants [2], in addition to cry-based recognition of infant [3]. With respect to various diseases, birth asphyxia and other breathing-related conditions, such as Sudden Infant Death Syndrome (SIDS) are leading cause of death for newborns [4]. Clinical diagnosis of asphyxia requires analysis of an arterial blood sample of newborns to measure blood gasses, pH, oxygen saturation, and electrolytes which requires a blood gas - a routine procedure in developed countries, however, in many developing countries it is not, as this procedure is costly and logistics heavy. Hence, asphyxia is generally detected only from emergency and visual symptoms, such as pale and bluish limbs, however, by then severe neurological damage would have already been occured to the newborn [4, 5]. Similarly, acoustic characteristics of deaf infants depends on hearing loss, type and period of rehabilitation, and the age of pathology identification [6]. Thus, there is a need to develop a cost effective and non-invasive cry diagnostic tool to assist pediatrics to detect early warning signs of such pathologies. To that effect, this study proposes signal processing-based approach for infant cry classification task, where asphyxia and deaf cry samples are considered as pathological samples.

The earlier investigations on infant cry started around 1960s, where four types of infant cries such as pain, hunger, birth, and pleasure were identified [7]. Xie *et. al.* have identified ten distinct cry modes-reflecting pattern of variations of fundamental frequency or pitch (F_0) and its harmonics from the narrowband spectrogram [8]. This study, however, was done for analysis of normal infant cries and was extended in [1], where dysphonation and hyperphonation, was found to be correlated with pathological infant cries [1]. In these studies, narrowband spectrograms (having horizontal striations indicate F_0 and its harmonics) were used, where formant structures are barely visible due to quasi-periodic sampling of vocal tract spectrum via high pitch-source harmonics.

State-of-the-art cepstral features, such as Mel Frequency Cepstral Coefficients (MFCC) are also used recently for cry classification task using Gaussian Mixture Model (GMM) as classifiers [9], [10]. However, in the context of Heisenberg's uncertainity principle in signal processing framework, Short-Time Fourier Transform (STFT) which is used in MFCC has fixed time-frequency resolution in entire time-frequency plane. In addition, it fails to preserve *form-invariance* property, as analysis window used in STFT is function of *only*

time parameter [11]

To that effect, we propose a new method of feature extraction based on Constant-Q Transform (CQT) and its cepstral representation, i.e., Constant Q Cepstral Coefficients (CQCC), which was originally puposed in antispoofing literature [12]. The feature extraction procedure of CQT is such that it has large number of bins located at the low frequency regions, where F_0 and its lower harmonics are present that are reasonably sufficient to analyze the characteristics of infant cry. Hence, CQT is able to preserve the pitch information as compared to the narrowband undersampled spectrogram derived from STFT. The CQT has variable spectro-temporal resolution in time-frequency plane, i.e., the analysis window function used in CQT is function of both time and frequency parameters and hence, it helps to achieve the form-invariance property-a desirable attribute of feature descriptors for pattern classification in the spectral-domain, which is not possible to achieve by the traditional STFT. Thus, the cry modes from the tradiational scpectrogram will not obey form-invariance property in spectral-domain and hence, may not be effective for infant cry classification task. Moreover, Brown's original investigations on CQT were motivated for improving resolution of notes in western music [13]. Perception, and memorising of melody and rhythm (i.e., prosody) start about third trimester of gestation, infants have remarkable musical predisposition, where melody contour (i.e., F_0 and its dynamics) is most salient for them [14]. Hence, we propose to employ CQT-based feature extraction in order to capture melodic structure in cry via F_0 and its dynamics for infant cry classification. To that effect, infant cry classification experiments are performed in this paper using CQCC with widely used GMM and SVM classifiers.

2. THE CONSTANT-Q TRANSFORM

The Discrete Fourier Transform (DFT) is nothing but uniform sampled version of Discrete-Time Fourier Transform (DTFT) performed on each frame of the speech signal [13]. Let $x(n)$ be the discrete-time input speech signal having sampling frequency, F_s. Then, STFT of $x(n)$ is given by [15]:

$$X(\omega,\tau) = \sum_{n=-\infty}^{\infty} x(n)\cdot \mathsf{w}(n,\tau)\cdot e^{-j\omega n}, \qquad (1)$$

where $\mathsf{w}(n,\tau)$ represents the analysis window, centered at time τ. It should be noted that $\mathsf{w}(n,\tau)$ is function of *only* time parameter τ as independent variable. Furthermore, let $y(n)$ represents a frame of the speech signal, then the DFT, $Y(k)$, of the $y(n)$ can be represented as:

$$Y(k) = \sum_{n=0}^{N-1} y(n)\cdot e^{-j(\frac{2\pi}{N})kn}, \qquad (2)$$

where k is the frequency bin index, and $\omega_{DFT} = (2\pi k)/N$. In this work, we have employed CQCC instead of STFT-based feature sets. CQT provides high frequency resolution in low frequency regions. In CQT, the quality factor Q of the subband filters in the filterbank remains constant and hence, frequency bins are geometrically-spaced as in Brown's original work [13]. The CQT of a signal $y(n)$ is represented as:

$$Y^{CQT}(k) = \frac{1}{N(k)} \sum_{k=0}^{N(k)-1} y(n)\mathsf{w}(n,k)e^{-j\left(\frac{2\pi}{N(k)}Qn\right)}, \qquad (3)$$

where $\omega_{CQT} = (2\pi Qn)/N(k)$, and $\mathsf{w}(n,k)$ is analysis window which has the identical shape for analysis of each frequency component f_k, however, its length is determined by $N(k)$ and thus, it is function of both n and k, where $N(k) = Q(F_s/f_k)$. Table 1 shows the window length for the set of parameters of CQT for infant cry classification. It can be observed from Table 1 that the window length varies w.r.t. the f_k, and it reduces with increase in f_k. Window length is very high for lower frequency regions, which provides the high frequency resolution and hence, the infant cry characteristics in low frequency regions can be effectively captured by the CQT. Since the quality factor Q is the ratio of center frequency to the bandwidth, it is given by [13]:

$$Q = \frac{f_k}{\Delta f_k} = \frac{f_k}{f_{k+1}-f_k} = \frac{1}{2^{1/B}-1}, \qquad (4)$$

where B represents the number of bins per octave, and f_k represents the frequency of k^{th} spectral component which is given as:

$$f_k = (2^{(k-1)/B})f_{min}, \qquad (5)$$

where f_{min} is the minimum frequency of the signal. Furthermore, we resampled the magnitude spectrum of CQT to linear scale in order to reduce the number of frequency bins in the feature set [12]. Furthermore, we have converted the geometrically-spaced frequency scale to linearly-spaced in order to preserve the orthogonality of the discrete cosine transform (DCT). In CQT, frequency bins are geometrically-spaced and hence, the signal reconstruction can be assumed as a downsampling operation over the initial k bins, which corresponds to low frequency and as an upsampling operation for the left over K-k bins, which corresponds to high frequency. The details of the resampling can be found in [12]. Performing DCT on resampled CQT gives CQCC feature set.

Table 1. Window length in samples as a function of analysis frequency (f_k). After [13].

k	Frequency (Hz)	# Samples	Duration (in ms)
1	100	29547	1340
100	204.37	14457	655.64
200	420	7022	318.48
400	1783	1657	75.15
600	7556	391	17.73

2.1. Form-Invariance Property of CQT

For the sake of simplicity, we consider continous-time version of FT, STFT, and CQT. If $x(n)$ and $X(\omega)$ are Fourier transform-pair, then time-scaling property of Fourier Transform implies [11], [15]:

$$\mathcal{F}\{x(\alpha t)\} = \frac{1}{|\alpha|} X\left(\frac{\omega}{\alpha}\right), \tag{6}$$

and thus, a linear time-scaling corresponds to frequency scaling by an *inverse* factor of $\frac{1}{a}$ and vice-versa indicating the *form* of spectrogram is unaffected and hence, the name form invariance. However, this property does not hold for the traditional STFT, where analysis window function is dependent *only* on time parameter. In particular, Schroeder and Atal defined the STFT through a practically realizable bandpass filters [16]. In particular,

$$F(t,\omega) = \int_{-\infty}^{t} f(\tau)\mathsf{w}(t-\tau)e^{-j\omega\tau}d\tau. \tag{7}$$

For form-invariance of STFT, we must have

$$F(t,\omega) = \gamma F(\alpha t, \beta\omega), \tag{8}$$

where α and β are scaling factor for time and frequency, respectively. However, it is shown in the literature that realization of eq. (8) yields the necessary and sufficient condition on weighting (i.e., window) function which belongs to the class of single term power functions, i.e., $\mathsf{w}(t) = a.t^b, t > 0$, and as per stability condition for Linear-Time Invariant (LTI) filter, this filter is unstable and hence, practically not realizable. However, it is interesting to note that if the window function is made to be frequency-dependent, i.e., $\mathsf{w}(t) \equiv \omega(t,\omega)$ (as in the case of CQT). In particular, equation (7) becomes

$$F(t,\omega) = \int_{-\infty}^{t} f(\tau)\mathsf{w}(t-\tau,\omega)e^{-j\omega\tau}d\tau, \tag{9}$$

the form-invariance property, i.e, eq. (9) is satisfied by eq. (10) for the window function, i.e.,

$$\mathsf{w}(t,\omega) = v(t,\omega)t^b \qquad t>0, \quad k>0, \quad \omega>0, \tag{10}$$

where v(tω) is an arbitrary real function of (tω), and b is real constant and function $\mathsf{w}(t,\omega)$ also satisfy stability condition for LTI filter, i.e.,

$$\int_{-\infty}^{\infty} |\mathsf{w}(t,\omega)|dt < \infty. \tag{11}$$

Furthermore, equation (10) also holds for window function considered in most practical model and short-time analysis performed by peripheral auditory system. For example, original model developed by Flanagan [17] represents the window function for the mechanical spectral analysis due to the movements of basilar membrane in the cochlea of human ear [11]. In particular, $\mathsf{w}(t,\omega) = (t\omega)^2 e^{\frac{-t\omega}{2}}$ which is similar to equation (10).

3. EXPERIMENTAL SETUP

Baby Chillanto databaseis used in this work. The details of the dataset can be studied in [18, 19]. Experiments are performed using 10-fold cross-validation.

Proposed CQCC feature set is employed with *90*-dimensions (90-*D*), which includes static, Δ, and $\Delta\Delta$ features. For fair comparison, we also extracted the *90-D* feature sets from STFT, named as *cepstrals*. Furthermore, we used the state-of-the-art MFCC feature set, extracted from the magnitude spectrum along with Mel filterbank that uses Mel-scaled

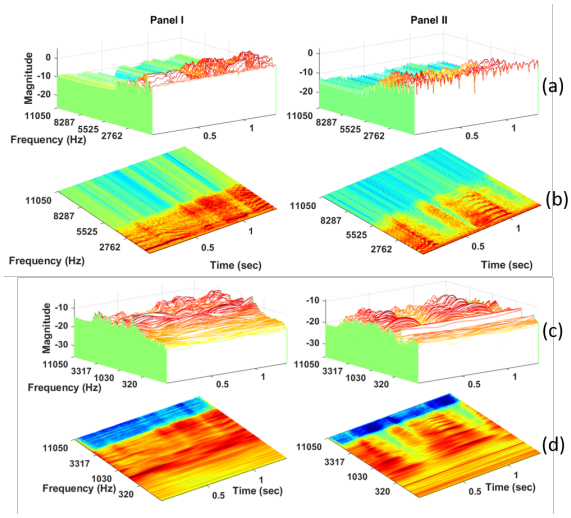

Fig. 1. Panel-I and Panel-II depicts the spectrographic analysis for healthy (normal) and pathology (asphyxia) infant cry signal: (a) the waterfall plot for STFT, (b) the top view of the STFT waterfall plot, (c) waterfall plot for CQT, and (d) the top view of the CQT waterfall plot.

bandpass filters [20]. In LFCC, Mel-scaled bandpass filters are replaced by linearly-spaced bandpass filters. For, LFCC and MFCC, we preserved initial *13*-dimensions (13-*D*), and then Δ and $\Delta\Delta$ coefficients are appended to it, which makes *39-D* feature sets. In this paper, we use two state-of-art classifiers, GMM and SVM, which are commonly used for infant cry classification task [1, 9]. The SVM was utilized in [21] and hence, we employed it as a classifier for another baseline architecture. The details of GMM and SVM can be studied in [22]. Furthermore, performance of various systems is evaluated using two performance metrics, namely, % classification accuracy and % Equal Error Rate (EER) [23].

4. EXPERIMENTAL RESULTS

4.1. Spectrographic Analysis

Panel-I and Panel-II shows the waterfall plots and corresponding top view of STFT and CQT for healthy *vs.* pathology cry signal, respectively. Fig. 1(a) and Fig. 1(b) shows the waterfall plot of STFT and its top view, where it can be observed that F_0 of the normal signal occurs above 300 Hz. Whereas, for pathological cry signal, the anomaly in the cry signal, which appears like F_0, is estimated in lower frequency regions. From Fig. 1(c) and Fig. 1(d), it can be observed that CQT emphasizes this anomaly in a much better way due to its high frequency resolution for lower frequencies.

4.2. Results

All the experiments in this paper are performed using *10*-fold cross validation. Initially, we performed the experiments by varying the f_{min} in eq. (5) with Hanning window in CQT and keeping the 512 number of Gaussian mixtures. It can be

Table 2. Results in % classification accuracy (Acc) for various f_{min} (Hz) of using GMM.

f_{min}	Acc.	f_{min}	Acc.	f_{min}	Acc.	f_{min}	Acc.
5	98.7	10	99.4	20	98.2	50	99.1
100	**99.8**	150	98.8	200	98.6	250	98.9

observed from Table 2 that best possible results are obtained with f_{min} = 100 Hz. This might be due to the fact that infant cry generally consists of F_0 above 250 Hz and hence, f_{min} = 100 Hz would be the optimum choice to capture the anomaly in the infant cry signal. Furthermore, experiments are performed w.r.t. various analysis window by keeping the 512 number of Gaussian mixtures in GMM as shown in Table 3. It can be observed that relatively better results are obtained using Hanning window. Furthermore, we also analyzed the performance w.r.t. number of Gaussian mixtures in GMM, and it is observed from Table 4 that 512 Gaussian mixtures are more suitable to estimate distribution of this data.

Table 3. Results (in % classification accuracy) for various window functions using GMM.

Window	Acc.	Window	Acc.
Hanning	**99.82**	Hamming	99.60
Gaussian	98.81	Rectangular	97.75

The experimental results obtained (in % classification accuracy and % EER) using combination of the various feature sets along with the GMM and SVM classifiers are reported in Table 5. It can be observed that relatively better performance is obtained for the proposed CQCC feature set using both GMM and SVM classifiers. Furthermore, it can be observed that CQCC and MFCC performs better than the *cepstrals* and LFCC, respectively. Here, MFCC and CQCC feature sets are designed w.r.t. perception of sounds in human auditory systems, which uses non-linear (in particular, logarithmic) scale along frequency-axis. Hence, we can conclude that the human auditory system-based features performing better as compared to linear-scale features for the pathology cry detection. The similar trends in results, as that of in

Table 4. Results (in % classification accuracy) w.r.t. number of mixtures.

Mixtures	64	128	256	512	1024
Accuracy	97.53	99.43	98.94	**99.82**	98.67

Table 5, are observed in DET curves (having discontinuities due to less genuine and imposter trials because of insufficient data) shown in Fig. 2. Furthermore, the performance of the proposed feature set is also validated by performing the standard statistical testing. To that effect, we have performed the *10*-fold cross-validation experiment for 50 times for each feature set and it was observed that the mean and median values of % classification accuracy for CQCC feature set are better than MFCC and LFCC feature sets, indicating statistical significance of proposed CQCC feature set. On the whole, proposed CQCC feature set performs better than the existing features for various evaluation factors, may be due to presentation of *form-invariance* property and CQT so that CQCC as feature descriptors is able to represent discriminative features of normal *vs.* pathological infant cry.

Table 5. Results in (% classification accuracy and % EER) for various feature sets using GMM as a classifier.

		MFCC	LFCC	Cepstrals	CQCC
GMM	Acc.	98.55	98.28	98.68	**99.82**
	EER	1.23	0.50	0.47	**0.44**
SVM	Acc.	88.11	80.18	80.62	**91.19**
	EER	12.72	18.78	17.73	**6.38**

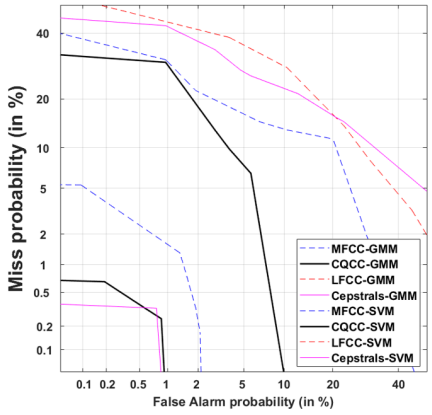

Fig. 2. DET plots for various features using GMM and SVM classifiers.

5. SUMMARY AND CONCLUSIONS

This study proposed CQCC feature set for infant cry classification task. Experiments performed with various evaluation factors, such as classifiers structures indicate better performance of CQCC than commonly used MFCC, LFCC, and STFT-based features. We believe that CQCC feature set preserve the *form-invariance* property thereby making feature descriptors invatiant w.r.t. linear scaling and hence, preserve discriminative features of normal *vs.* pathological infant cries. One of the limitation of this work could be to analyze effectiveness of CQCC on various deep learning classifiers. However, these models require large amount of training data which is difficult to collect for infant cry classification task, more so, for pathological infant cries. To alleviate this, future work can be to explore various data augmentation methods, such as pitch, time-scale, and tempo modification. It is less understood in the literature that whether form-invariance property is actually exploited for the perception of sounds, however, such property holds for peripheral auditory system and thus, linking these two aspects remain open research question which may have interesting technological applications in vocoders, anti-spoofing, etc.

6. REFERENCES

[1] Hemant A. Patil, "Cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry," in *A. Newtein (Ed.) Advances in Speech Recognition, Springer*, pp. 323–348. 2010.

[2] Joshua James Engelsma, Debayan Deb, Kai Cao, Anjoo Bhatnagar, Prem Sewak Sudhish, and Anil K. Jain, "Infant-id: Fingerprints for global good," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[3] Hemant A. Patil, "Infant identification from their cry," in *2009 Seventh International Conference on Advances in Pattern Recognition*. IEEE, 2009, pp. 107–110.

[4] Charles C. Onu, Innocent Udeogu, Eyenimi Ndiomu, Urbain Kengni, Doina Precup, Guilherme M Sant'Anna, Edward Alikor, and Peace Opara, "Ubenwa: Cry-based diagnosis of birth asphyxia," *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA*, 2017.

[5] Charles C. Onu, Jonathan Lebensold, William L. Hamilton, and Doina Precup, "Neural transfer learning for cry-based diagnosis of perinatal asphyxia," *ICLR Workshop*, 2019.

[6] Kathiresan Manickam and Haizhou Li, "Complexity analysis of normal and deaf infant cry acoustic waves," in *Fourth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2005.

[7] O Wasz-Höckert, T.J. Partanen, V. Vuorenkoski, K Michelsson, and E Valanne, "The identification of some specific meanings in infant vocalization," *Experientia*, vol. 20, no. 3, pp. 154–154, 1964.

[8] Qiaobing Xie, Rabab K. Ward, and Charles A Laszlo, "Automatic assessment of infants' levels-of-distress from the cry signals," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 253, 1996.

[9] Hesam Farsaie Alaie, Lina Abou-Abbas, and Chakib Tadj, "Cry-based infant pathology classification using gmms," *Speech Communication*, vol. 77, pp. 28–52, 2016.

[10] Chunyan Ji, Thosini Bamunu Mudiyanselage, Yutong Gao, and Yi Pan, "A review of infant cry analysis and classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–17, 2021.

[11] G. Gambardella, "Time scaling and short-time spectral analysis," *The Journal of the Acoustical Society of America (JASA)*, vol. 44, no. 6, pp. 1745–1747, 1968.

[12] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "Constant-Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language,2017 Bilbao, Spain*, vol. 45, pp. 516–535, Bilbao, Spain, June 21-24, 2017.

[13] Judith C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America (JASA)*, vol. 89, no. 1, pp. 425–434, 1991.

[14] Lotte Armbrüster and et. al., "Musical intervals in infants' spontaneous crying over the first 4 months of life," *Folia Phoniatrica et Logopaedica*, vol. 73, no. 5, pp. 401–412, 2021.

[15] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1^{st} Edition, Pearson Education India, 2015.

[16] Manfred R. Schroeder and Bishnu S. Atal, "Generalized short-time power spectra and autocorrelation functions," *The Journal of the Acoustical Society of America (JASA)*, vol. 34, no. 11, pp. 1679–1683, 1962.

[17] James L. Flanagan, *Speech analysis synthesis and perception*, vol. 3, Springer Science & Business Media, 2013.

[18] Hardik B Sailor, Madhu R Kamble, and Hemant A Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 666–670.

[19] Alejandro Rosales-Pérez, Carlos A Reyes-García, Jesus A Gonzalez, Orion F Reyes-Galaviz, Hugo Jair Escalante, and Silvia Orlandi, "Classifying infant cry patterns by the genetic selection of a fuzzy model," *Biomedical Signal Processing and Control*, vol. 17, pp. 38–46, 2015.

[20] Steven Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[21] Kosuke Akimoto, Seng Pei Liew, Sakiko Mishima, Ryo Mizushima, and Kong Aik Lee, "POCO: A voice spoofing and liveness detection corpus based on pop noise," in *INTERSPEECH*, Shanghai, China, October 2020, pp. 1081–1085.

[22] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[23] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, 1997, pp. 1895–1898.