INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Comparison of GIF- and SSL-based features in pathological-voice detection

Akira Sasou[1], *Yang Chen*[1]

[1]National institute of advanced industrial science and technology, AIST
a-sasou@aist.go.jp, chenyang.chin@aist.go.jp

Abstract

A system that automatically detects voice pathology from acoustic signals enables non-invasive, low cost, and objective assessment of speech disorders. Therefore, it is expected to accelerate and improve the diagnosis and clinical treatment of patients. Pathological voices are symptoms of impairments in the articulation of speech sound, fluency, and/or voice. We consider that direct extraction of features from the glottal flow estimated by glottal inverse filtering (GIF) is a promising approach to pathological-voice detection. To precisely estimate the glottal flow, we propose a novel GIF method that combines constrained autoregressive hidden Markov model (CAR–HMM) analysis with automatic topology generation of the excitation HMM. To evaluate the effectiveness of the features extracted from the estimated glottal flow during pathological-voice detection, we employ the Saarbrücken Voice Database. We also compare the features obtained by the proposed CAR–HMM with those obtained by pre-trained models based on self-supervised learning (SSL). The experimental results confirmed that the CAR–HMM-based method can outperform the SSL-based methods.

Index Terms: pathology voice, glottal flow, auto-regressive hidden Markov model, self-supervised learning

1. Introduction

Speech conveys rich information including not only linguistic content but also the identity, gender, age, emotions, and health state of the speaker. Recently, systems that automatically detect voice pathology from acoustic signals have attracted much interest [2,3]. Especially, by providing non-invasive, low cost, objective assessments of disorders, they are expected to accelerate and improve the diagnosis and clinical treatment of patients [1].

Speech production requires the cooperation of multiple organs: (1) the nervous system, (2) respiratory system, and (3) the vocal cords and vocal tracts [7]. Speech disorders result from infectious, physiological, or psychogenic disruptions to any of these systems. Professionals who use their voice excessively at work, such as singers and teachers, are especially vulnerable to such disruptions [8,9]. Pathological voices are symptoms of impaired articulation of speech sound, fluency, and/or voice [10]. Therefore, by separately extracting the vocal cord- and vocal tract-related acoustical features from speech signals, we should provide important data for the acoustical assessment and identification of speech disorders. Although methods for estimating and parameterizing the vocal tract system are well established, methods for estimating the glottal flow generated by the vibrations of vocal cords appear to be under-investigated [11]. The effectiveness of glottal flow-related features in the analysis and detection of voice pathologies has only recently been reported [4,5,12,13].

Methods that estimate the glottal flow from voiced speech signals are referred to as glottal inverse filtering (GIF) methods. Speech signals can be measured only as an output of a composite system including the glottal flow, vocal tract, and lip radiation. As the glottal flow cannot be directly measured from speech signals using non-invasive measuring techniques, we cannot easily acquire the ground truth of glottal flow signals corresponding to the acquired speech signals; moreover, GIF is hardly realized through supervised deep learning. Several GIF methods have been handcrafted based on speech-signal-analyses and knowledge of the speech production mechanism [15, 16, 17, 18].

To improve the estimation accuracy of vocal-tract characteristics, Sasou et al. [19,20] proposed a speech analysis method based on an autoregressive hidden Markov model (AR–HMM), which is especially effective on high fundamental-frequency speech signals. The HMM was introduced as an excitation-source model and the states were concatenated in a ring topology to circulate the state transition, representing the periodicity of voiced speech. However, the prediction residual obtained from inverse filtering through a learned AR filter, and the learned HMM to which the generated prediction-residual conforms, might lack sufficient information related to a physically observable signal. For this reason, AR–HMM-based analysis is not directly usable as a GIF method. To address this problem, Sasou [14] recently proposed a constrained AR–HMM (CAR–HMM) analysis that directly models the glottal flow derivative under constraints on the AR filter estimation, where the HMM of a ring topology is adopted. When applying CAR–HMM to the analysis of pathological voices, particularly those caused by vocal-cord disorders, the ring-topology assumption of the HMM might need to be relaxed to emphasize the irregularities in glottal flow. Previously, Sasou [21,34] proposed a successive state splitting (SSS) method that can automatically generate the topology of the excitation-source HMM for the AR–HMM-based speech analysis method.

In the present paper, we propose a novel GIF method that combines CAR–HMM with automatic topology generation of the excitation-source HMM. This approach is expected to more precisely estimate the glottal flow than CAR–HMM with a ring topology. We evaluate the effectiveness of the features extracted from the estimated glottal flow during pathology voice detection. Meanwhile, self-supervised learning (SSL) promises to realize a single universal model for a wide variety of tasks and domains [22]. SSL is classified as an unsupervised learning approach because it attempts to discover the naturally occurring patterns in training samples, which cannot be preassigned with labels or scores [23]. We also compare the proposed CAR–HMM-derived features with those obtained using pre-trained SSL-based models [24].

10.21437/Interspeech.2023-1418

2. Proposed GIF method

2.1. Parameter estimation algorithm for CAR–HMM

This subsection briefly reviews the iterative parameter estimation (IPE) algorithm of CAR–HMM (see [14] for details). In the following derivation, P denotes the order of the vocal-tract AR filter and $a_k^{(i)}$ $(k = 1, \cdots, P)$ denotes the AR coefficients obtained in the i-th iteration. The AR filter takes the following form:

$$V^{(i)}(z) = \frac{1}{A^{(i)}(z)} = \frac{1}{1-\sum_{k=1}^{P} a_k^{(i)} z^k} \tag{1}$$

where $A^{(i)}(z)$ denotes the inverse filter. The AR coefficient vector is denoted as

$$\mathbf{a}^{(i)} = \begin{bmatrix} a_1^{(i)} & a_2^{(i)} & \cdots & a_P^{(i)} \end{bmatrix}^T \in \mathbb{R}^P \tag{2}$$

Let N represent the number of speech-signal samples in the analysis frame. The speech-signal samples $x_n (n = P, \cdots, N-1)$ can be described by the following vector:

$$\mathbf{x}_P = [x_P \quad x_{P+1} \quad \cdots \quad x_{N-1}]^{\mathrm{T}} \in \mathbb{R}^{N-P}. \tag{3}$$

Let $\mathbf{\Omega}$ be the following matrix:

$$\mathbf{\Omega} = [\mathbf{x}_{P-1} \quad \mathbf{x}_{P-2} \quad \cdots \quad \mathbf{x}_0] \in \mathbb{R}^{(N-P)\times P}. \tag{4}$$

The predicted residual samples $\tilde{e}_n^{(i)}, (n = P, \cdots, N-1)$ can be described by the following vector:

$$\tilde{\mathbf{e}}_P^{(i)} = \begin{bmatrix} \tilde{e}_P^{(i)} & \tilde{e}_{P+1}^{(i)} & \cdots & \tilde{e}_{N-1}^{(i)} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{N-P}. \tag{5}$$

The random variables e_n $(n = P, \cdots, N-1)$ in the glottal flow derivative are assembled into the following vector:

$$\mathbf{e}_P = [e_P \quad e_{P+1} \quad \cdots \quad e_{N-1}]^{\mathrm{T}} \in \mathbb{R}^{N-P}. \tag{6}$$

This random vector of the glottal flow derivative is assumed to follow a multidimensional normal distribution:

$$\mathbf{e}_P \sim \mathrm{N}\left(\mathbf{m}_P^{(i)}, \mathbf{\Sigma}_P^{(i)}\right), \tag{7}$$

where the i-th estimates of the expectation vector and the covariance matrix are respectively given by

$$\mathbf{m}_P^{(i)} = \begin{bmatrix} m_P^{(i)} & m_{P+1}^{(i)} & \cdots & m_{N-1}^{(i)} \end{bmatrix}^T \in \mathbb{R}^{N-P}, \tag{8}$$

$$\mathbf{\Sigma}_P^{(i)} = \mathrm{diag}\left(v_P^{(i)}, v_{P+1}^{(i)}, \cdots, v_{N-1}^{(i)}\right) \in \mathbb{R}^{(N-P)\times(N-P)}. \tag{9}$$

The IPE algorithm is executed as follows:

[IPE-Step 1] Set the initial population parameters of the glottal flow derivative random vector to the following values:

$$\mathbf{m}_P^{(0)} = \mathbf{0}, \tag{10}$$

$$\mathbf{\Sigma}_P^{(0)} = \mathrm{diag}\left(v_P^{(0)}, v_{P+1}^{(0)}, \cdots, v_{N-1}^{(0)}\right). \tag{11}$$

Here, the variances are initially set to positive random values or unity. The next step begins with $i = 0$.

[IPE-Step 2] Adjust the AR coefficients such that the prediction-residual vector becomes the realization vector of the glottal flow derivative random vector conforming to $\mathrm{N}\left(\mathbf{m}_P^{(i)}, \mathbf{\Sigma}_P^{(i)}\right)$ and the gains of the inverse filter are constrained as specified. The constraints on the DC and Nyquist frequency gains are given by:

$$\begin{aligned} &A^{(i)}\left(e^{j0}\right) = 1 - \textstyle\sum_{k=1}^{P} a_k^{(i)} = 1 - \mathbf{c}^{\mathrm{T}}\mathbf{a}^{(i)} = l_{dc}, \\ &\mathbf{c} = [1 \quad 1 \quad \cdots \quad 1]^{\mathrm{T}} \in \mathbb{R}^P, \\ &A^{(i)}\left(e^{j\pi}\right) = 1 - \textstyle\sum_{k=1}^{P} (-1)^k a_k^{(i)} = 1 - \mathbf{d}^{\mathrm{T}}\mathbf{a}^{(i)} = l_{nq}, \\ &\mathbf{d} = [(-1)^1 \quad (-1)^2 \quad \cdots \quad (-1)^P]^{\mathrm{T}} \in \mathbb{R}^P. \end{aligned} \tag{12}$$

We must now solve the following optimization problem:

$$\mathbf{a}^{(i+1)} = \underset{\mathbf{a}}{\mathrm{argmax}}\, \mathrm{L}\left(\tilde{\mathbf{e}}_P(\mathbf{a}); \mathbf{m}_P^{(i)}, \mathbf{\Sigma}_P^{(i)}\right),$$

$$\text{subject to } 1 - \mathbf{c}^{\mathrm{T}}\mathbf{a} = l_{dc} \text{ and } 1 - \mathbf{d}^{\mathrm{T}}\mathbf{a} = l_{nq}, \tag{13}$$

Here, the prediction-residual vector is a function of the AR coefficient vector as

$$\tilde{\mathbf{e}}_P(\mathbf{a}) = \mathbf{x}_P - \mathbf{\Omega}\mathbf{a}. \tag{14}$$

The solution that maximizes the objective function is given as follows, where (i) is omitted to simplify the notations:

$$\begin{aligned} &\mathbf{a}^{(i+1)} = \mathbf{a}_0 - \lambda\mathbf{G}\mathbf{c} - \gamma\mathbf{G}\mathbf{d}, \\ &\mathbf{a}_0 = [\mathbf{\Omega}^{\mathrm{T}}\mathbf{\Sigma}_P^{-1}\mathbf{\Omega}]^{-1}\mathbf{\Omega}^{\mathrm{T}}\mathbf{\Sigma}_P^{-1}(\mathbf{x}_P - \mathbf{m}_P), \\ &\mathbf{G} = (\mathbf{\Omega}^{\mathrm{T}}\mathbf{\Sigma}_P^{-1}\mathbf{\Omega})^{-1}, \\ &\begin{bmatrix} \lambda \\ \gamma \end{bmatrix} = \frac{1}{(\mathbf{c}^{\mathrm{T}}\mathbf{G}\mathbf{c})(\mathbf{d}^{\mathrm{T}}\mathbf{G}\mathbf{d}) - (\mathbf{c}^{\mathrm{T}}\mathbf{G}\mathbf{d})(\mathbf{d}^{\mathrm{T}}\mathbf{G}\mathbf{c})} \times \\ &\begin{bmatrix} \mathbf{d}^{\mathrm{T}}\mathbf{G}\mathbf{d} & -\mathbf{c}^{\mathrm{T}}\mathbf{G}\mathbf{d} \\ -\mathbf{d}^{\mathrm{T}}\mathbf{G}\mathbf{c} & \mathbf{c}^{\mathrm{T}}\mathbf{G}\mathbf{c} \end{bmatrix} \begin{bmatrix} \mathbf{c}^{\mathrm{T}}\mathbf{a}_0 + l_{dc} - 1 \\ \mathbf{d}^{\mathrm{T}}\mathbf{a}_0 + l_{nq} - 1 \end{bmatrix}. \end{aligned} \tag{15}$$

Using (14), the prediction-residual vector is updated as

$$\tilde{\mathbf{e}}_P^{(i+1)} = \mathbf{x}_P - \mathbf{\Omega}\mathbf{a}^{(i+1)}. \tag{16}$$

[IPE-Step 3] Check the convergence status. If the following inequality is satisfied for small ε, or if the number of iterations reaches the pre-specified value, the iterations are terminated:

$$\frac{\left|\log\left\{\mathrm{L}\left(\tilde{\mathbf{e}}_P^{(i+1)}; \mathbf{m}_P^{(i)}, \mathbf{\Sigma}_P^{(i)}\right)\right\} - \log\left\{\mathrm{L}\left(\tilde{\mathbf{e}}_P^{(i)}; \mathbf{m}_P^{(i)}, \mathbf{\Sigma}_P^{(i)}\right)\right\}\right|}{\left|\log\left\{\mathrm{L}\left(\tilde{\mathbf{e}}_P^{(i)}; \mathbf{m}_P^{(i)}, \mathbf{\Sigma}_P^{(i)}\right)\right\}\right|} < \varepsilon. \tag{17}$$

Otherwise, proceed to IPE-Step 4.

[IPE-Step 4] Update the population parameters of the glottal-flow-derivative random vector by maximizing the likelihood as follows:

$$\mathbf{m}_P^{(i+1)}, \mathbf{\Sigma}_P^{(i+1)} = \underset{\mathbf{m}_P, \mathbf{\Sigma}_P}{\mathrm{arg\,max}}\, \mathrm{L}\left(\tilde{\mathbf{e}}_P^{(i+1)}; \mathbf{m}_P, \mathbf{\Sigma}_P\right), \tag{18}$$

where the updated prediction-residual vector is regarded as the newly realized random vector of the glottal flow derivative. The likelihood can be maximized through the following procedure. The HMM has S states, each with a unique number assigned from $\mathbb{S} = \{1, \cdots, S\}$. The HMM is learned using the updated prediction-residual time sequence of the Baum–Welch algorithm, which obtains the population parameters $\mu_s^{(i+1)}$ and $\sigma_s^{2,(i+1)}$ $(s \in \mathbb{S})$ of each output PDF. The Viterbi algorithm then finds the most likely state-transition sequence $s_n^{(i+1)} \in \mathbb{S}$ $(n = P, \cdots, N-1)$ corresponding to the prediction-residual time sequence. The random variable e_n of the glottal flow derivative is assumed to conform to the following population parameters: expectation $m_n^{(i+1)} = \mu_{s_n}^{(i+1)}$ and variance $v_n^{(i+1)} = \sigma_{s_n}^{2,(i+1)}$, where $(i+1)$ is omitted from $s_n^{(i+1)}$. The updated population parameters can be respectively described in vector and matrix forms as

$$\mathbf{m}_P^{(i+1)} = \begin{bmatrix} m_P^{(i+1)} & m_{P+1}^{(i+1)} & \cdots & m_{N-1}^{(i+1)} \end{bmatrix}^T, \tag{19}$$

$$\mathbf{\Sigma}_P^{(i+1)} = \mathrm{diag}\left(v_P^{(i+1)}, v_{P+1}^{(i+1)}, \cdots, v_{N-1}^{(i+1)}\right). \tag{20}$$

After setting i to i+1, the procedure returns to IPE-Step 2 and begins the next iteration.

2.2. CAR–HMM combined with successive state splitting

The SSS algorithm was originally proposed for optimizing a network of HMM states to an individual speaker [25] and was later expanded to the MDL–SSS algorithm, which conducts both contextual and temporal splitting with the MDL criteria as the splitting and stop criteria [26]. Applying MDL–SSS, we automatically generated the topology of the excitation-source HMM for AR–HMM-based speech analysis [21]. In the present paper, we combine the parameter estimation algorithm of CAR–HMM with MDL–SSS for precise modeling of the glottal flow derivative.

The algorithm first adopts a topology with a single state. In the following, $\phi(N_s)$ represents the AR–HMM parameter set comprising the AR coefficients and the parameters of the excitation-source HMM with N_s states. The AR–HMM

parameters $\phi(1)$ are then estimated using the CAR–HMM parameter estimation algorithm. Next, the $\phi(2)$ parameters are estimated with the two states concatenated in a ring state. These two models are compared under the MDL criterion. If the MDL of $\phi(N_s)$ exceeds the MDL of $\phi(N_s-1)$, then $\phi(N_s-1)$ is selected as the final model and the state-splitting algorithm finishes. Otherwise, the likelihood of each state is evaluated as

$$l(s)=\prod_{n\in\left\{n\middle|\begin{matrix}n=P,\cdots,N-1\\ s_n^{(i)}=s\end{matrix}\right\}}\frac{1}{\sqrt{2\pi\sigma_s^{2,(i)}}}\exp\left\{-\frac{\left(\hat{e}_n^{(i)}-\mu_s^{(i)}\right)^2}{2\sigma_s^{2,(i)}}\right\}. \quad (21)$$

The minimum likelihood state selected by $s^*=\mathrm{argmin}_{s\in\mathbb{S}}l(s)$ is then split in both the temporal and contextual directions, as depicted in Figure 1. Here ϕ_t and ϕ_c represent the parameter sets of CAR–HMM split in the temporal and contextual directions, respectively. After evaluating the MDLs of both CAR–HMMs, the parameter set of the CAR–HMM with the lowest MDL is adopted as $\phi(N_s+1)$. These processes are iterated until the stop condition is satisfied.

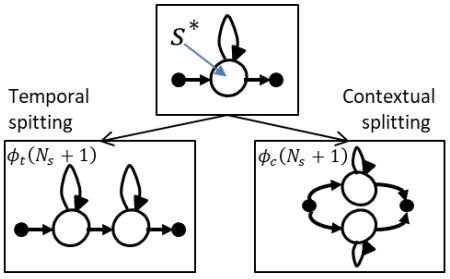

Figure 1: *State splitting in the temporal and contextual directions.*

3. Experiments

3.1. Database

The following experiments were performed on the Saarbrücken Voice Database (SVD) [27], which contains recordings of 1002 speakers with a wide range of voice disorders (454 male and 548 female). Each session contains the recordings of /a/, /i/, and /u/ vowels uttered with seven kinds of pitch patterns, along with short phrases. The SVD uses 71 different pathology labels. One thousand and ninety three pathological recording sessions are assigned a single label and 263 pathological recording sessions are assigned multiple labels [6]. We counted the number of appearance frequencies of each of the 71 labels and selected the 15 labels that appeared at least 30 times. In the following experiments, we used 1213 pathological recording sessions assigned one of the 15 labels along with 685 control recording sessions. To balance the number of female and male speakers' recordings, we first separated the recordings in each label class by gender. We then partitioned the recordings in the respective label class and respective gender class into training, validation, and test datasets at a ratio of 8:1:1. We combined the female and male speakers' recordings in the respective label class and the respective dataset to generate gender-balanced datasets. The experimental task was pathological-voice detection, where the vowel recordings were classified as either pathological or normal. For this purpose, we replaced all 15 labels of pathological vowel utterances with "pathology" and labeled all control vowels as "normal."

3.2. CAR–HMM-based feature extraction

In the following experiments, we used the recordings of the /a/ vowel with normal pitch. The recordings in the SVD were digitized with sampling at 50 kHz and quantized with 16-bit resolution. The recordings were down-sampled to 16 kHz. The vowel sounds were pre-emphasized through a high-pass filter $(1-0.99z^{-1})$ and framed with a frame length and shift period of 200 ms and 100 ms, respectively. Applying the CAR–HMM with SSS to each frame, we estimated the glottal flow derivative setting the order of the AR filter to 16, the maximum number of successively split states to 5, the number of iterations in the IPE algorithm to 20, and the constraints on both gains to $l_{dc}=l_{nq}=1$ in (12). The frame number of the glottal flow derivatives that were estimated by the CAR-HMM from one recording of the vowel sound depends on the length of the recording.

Next, we extracted the features of the Mel Filter Bank (MelFB) from the respective frames of the glottal flow derivative. After pre-emphasizing the respective frame of the glottal flow derivative through a high-pass filter $(1-0.97z^{-1})$, we calculated the fast Fourier transforms (FFTs) using analysis frames with a length and shift period of 10 ms and 1.25 ms, respectively. Triangular windows were applied to the FFT amplitudes to generate 40-dimensional MelFBs. From the respective frame of the glottal flow derivative, we obtained 153 MelFBs.

3.3. Transformers for CAR–HMM-based features

We employed a standard Transformer encoder composed of a stack of identical layers followed by a classification head and a SoftMax layer. Each layer of Transformer has two sub-layers: a multi-head self-attention mechanism and a simple, position-wise, fully connected feed-forward network [28]. The 153-long input sequence of MelFBs extracted from each frame was directly fed to the first layer of the Transformer with no positional embeddings. Our preliminary experimental results indicated that adding positional embeddings to MelFBs degraded the recognition accuracy. To ensure that Transformer could output a prediction for every MelFB in the input sequence, every vector in the output sequence was independently fed to the classification head. We then aggregated the predictions of all MelFBs extracted from one recording. The final decision (whether the recording belongs to the "pathology" or "normal" class) was determined with simple max-voting.

We implemented twelve Transformers with different model structures and training conditions. For each Transformer, Table 1 lists the number of layers, number of multi-headers in the self-attention mechanism, number of nodes in the hidden layer of the feed-forward network, and the structure of the classification head. During training, a residual dropout with a rate of 0.5 was applied to the output of each Transformer sub-layer. The batch size was set to 8, 16, or 32 depending on the memory consumption of the GPU. The maximum epoch number was set to 2000. In each training epoch, Transformer was evaluated on the validation dataset. After iterating 2000 epochs, we selected the best-scoring model as the optimal model. For optimization, we employed the Adam optimizer [29] with a learning ratio of 1e-5 or 1e-4. Each Transformer was iteratively trained ten times under the same conditions using different random seeds.

Table 1: *Conditions of the Transformers used in the experiments*

Model ID	#Layers	#Heads	#Nodes	Classification Head	Drop Rate	Batch Size	Epoch	Learning Rate
1	1	3	2048	linear,Relu,linear	0.5	32	2000	1.0E-05
2	4	3	2048	linear,Relu,linear	0.5	32	2000	1.0E-05
3	4	3	4096	linear,Relu,linear	0.5	32	2000	1.0E-05
4	6	3	4096	linear	0.5	32	2000	1.0E-05
5	4	3	4096	linear,Relu,linear	0.5	32	2000	1.0E-04
6	6	3	2048	linear	0.5	32	2000	1.0E-04
7	5	3	2048	linear,Relu,linear	0.5	32	2000	1.0E-04
8	5	3	2048	linear,Relu,linear,Relu,linear	0.5	32	2000	1.0E-04
9	4	3	4096	linear,Relu,linear,Relu,linear	0.5	16	2000	1.0E-04
10	4	3	4096	linear,Relu,linear	0.5	8	2000	1.0E-04
11	4	3	4096	linear,Relu,linear,Relu,linear	0.5	8	2000	1.0E-04
12	3	3	2048	linear,Relu,linear,Relu,linear	0.5	8	2000	1.0E-04

3.4. Self-supervised learning-based method

The performance of CAR–HMM in pathological-voice detection was compared with those of pre-trained models with different SSLs as acoustic-feature extractors. As the SSLs, we employed the modified Contrastive Predictive Coding (CPC) [30], wav2vec2.0 Large [31], and the hidden-unit BERT (HuBERT) Large [32]. The feature dimensions of the modified CPC, wav2vec.20 Large, and HuBERT Large are 256, 1024, and 1024, respectively.

Figure 2 shows the model for classifying the SSL-based features. Originally proposed in [33] for speech-emotion recognition, the model consists of one Transformer block and two convolution blocks combined in parallel. Both convolution blocks have the same structure, as represented in the lower part of Figure 2. In the following, we describe only the optimal model of the respective SSL among all examined models. Although the sequence length of the extracted features depended on the recording length of the vowel sound, the length of the input-feature sequences was fixed at 512. If the extracted feature sequence was longer than 512, it was truncated to the first 512 features. In contrast, if the feature sequence was shorter than 512, its length was extended by iteratively concatenating the original feature sequence. Like the Transformer used for CAR–HMM-based feature extraction, the Transformer block included only an encoder part. The number of encoder layers was set to one. The number of multi-headers of the self-attention mechanism was also set to one. The number of nodes in the hidden layer of the feed-forward network was set to 1024 for the modified CPC and 2048 for wav2vec.20 Large and HuBERT Large. The batch size was set to 16. In each training epoch, the Transformer was evaluated on the validation dataset. The best-scoring model was selected as the optimal model. For optimization, we used the Adam optimizer with a learning ratio of 1e-4. The parallel model was iteratively trained ten times under the same conditions for each SSL using different random seeds.

3.5. Results

The identification results on the validation and test datasets were evaluated in terms of their F1-scores. Tables 2 and 3 present the F1-scores of the CAR–HMM- and SSL-based methods, respectively. In each table, the maximum F1-score on each dataset is highlighted in bold type. On the validation dataset, the F1-scores of the SSL-based methods exceed those of the CAR–HMM-based methods; however, on the test dataset, the CAR–HMM-based methods tend to outperform the SSL-based methods. Therefore, the generalization ability of the CAR–HMM-based methods appear to be superior to the SSL-based methods. On both the validation and test datasets, the F1-scores of the CAR–HMM-based method were maximized in Model 10 (which achieved 75.15% and 75.27% on the validation and test datasets, respectively). Meanwhile, the SSL-based method achieved its maximum F1-scores with the modified CPC SSL (78.99% and 74.21% on the validation and test datasets, respectively). Comparing the F1-scores of the test dataset, we can conclude that the CAR–HMM-based method is more promising than the SSL-based method.

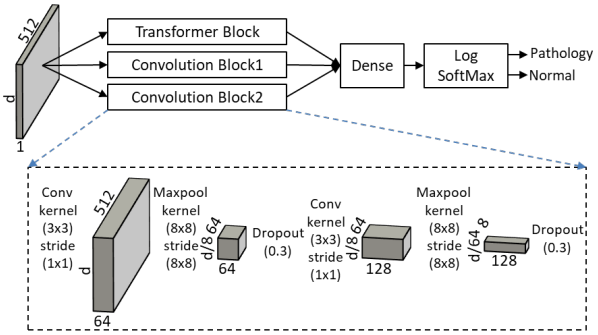

Figure 2: *Classification model for SSL-based features (*d *represents the dimension of the features).*

Table 2: *Results of the CAR–HMM-based methods.*

ModelID			1	2	3	4	5	6	7	8	9	10	11	12
F1-Score [%]	Validation	Avg	73.59	72.10	71.97	71.82	74.68	74.45	73.71	74.04	73.72	**75.15**	74.05	73.74
		SD	1.56	1.54	1.93	2.38	1.32	1.96	2.03	2.96	1.98	1.31	1.24	1.66
	Test	Avg	73.06	73.16	73.74	71.42	73.59	73.63	73.19	72.71	74.45	**75.27**	72.11	73.82
		SD	1.80	2.59	2.53	2.38	2.42	2.38	3.38	3.43	1.85	1.72	2.29	2.56

Table 3: *Results of the SSL-based methods.*

SSL ID			modified CPC	wav2vec2.0 Large	HuBERT Large
F1-Score [%]	Validation	Avg	**78.99**	77.98	78.02
		SD	1.36	1.79	2.73
	Test	Avg	**74.21**	71.08	71.06
		SD	1.49	1.78	2.58

4. Conclusions

We proposed a novel GIF method that combines CAR–HMM with automatic topology generation of the excitation-source HMM. The features obtained from the glottal flow derivative with the proposed GIF method were applied to pathological-voice detection. Comparison experiments showed that the CAR–HMM-based methods outperformed the SSL-based methods. Judging from these experimental results, the proposed GIF method is both valid and effective. One drawback of the proposed GIF method is the complicated algorithm, which is calculation-intensive and time-consuming. We are now planning a surrogate model of the proposed GIF method.

5. Acknowledgments

This paper is based on the results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

6. References

[1] D. Panek, A. Skalski, J.Gajda and R. Tadeusiewicz, “Acoustic analysis assessment in speech pathology detection,” Int. J. Appl. Math. Comput. Sci., 2015, Vol.25, No.3, pp.631-643.

[2] N.Saenz-Lechon, J.I.Godino-Llorente, V.Osma-Ruiz, P.Gomez-Vilda, “Methodological issue in the development of automatic

systems for voice pathology detection," Biomed. Signal Processing and Control, 1, pp.120-128, 2006.
[3] H. Wu, J. Soraghan, A. Lowit, and G. D. Caterina, "A Deep Learning Method for Pathological Voice Detection using Convolutional Deep Belief Network," in *Proceedings of INTERSPEECH 2018*, pp.446-450, Sep. 2018.
[4] S. R. Kadiri and P. Alku, "Analysis and Detection of Pathological Voice Using Glottal Source Features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367-379, Feb. 2020.
[5] Y. Wu, C. Zhou, Z. Fan, D. Wu, X. Zhang and Z. Tao, "Investigation and Evaluation of Glottal Flow Waveform for Voice Pathology Detection," in *IEEE Access*, vol. 9, pp. 30-44, 2021, doi: 10.1109/ACCESS.2020.3046767.
[6] M. Huckvale and C. Buciuleac, "Automated Detection of Voice Disorder in the Saarbrucken Voice Database: Effects of Pathology Subset and Audio Materials," *Proceedings of INTERSPEECH 2021*.
[7] D. Zhang and K. Wu, *Pathological Voice Analysis*. Springer Nature Singapore, 2020.
[8] A. E. Aronson, *Clinical Voice Disorders; An Interdisciplinary Approach*. New York: Theme Inc., 1985.
[9] N. R. William, "Occupational groups at risk of voice disorders: A review of the literature," *Occupational Med*., vol. 53, no. 7, pp. 456-460, 2003.
[10] J. A. Gomez-Garcia, L. Moro-Velazquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," Biomedical Signal Processing and Control, vol. 51, pp. 181-199, 2019.
[11] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Extraction and Utilization of Excitation Information of Speech: A Review," in Proceedings of the IEEE, vol. 109, no. 12, pp. 1920-1941, Dec. 2021, doi: 10.1109/JPROC.2021.3126493.
[12] N. P. Narendra and P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," Comput. Speech Lang., vol. 65, Jan. 2021, Art. no. 101117.
[13] V. M. Espinoza, M. Zañartu, J. H. Van Stan, D. D. Mehta, and R. E. Hillman, "Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction," J. Speech Lang., Hearing Res., vol. 60, no. 8, pp. 2159–2169, 2017.
[14] A. Sasou, "Glottal inverse filtering by combining a constrained LP and an HMM-based generative model of glottal flow derivative," Speech Communication, vol.104, 2018, pp.113-128, https://doi.org/10.1016/j.specom.2018.07.002.
[15] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 4, pp. 350-355, August 1979.
[16] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," Speech Communication, vol. 11, issues 2–3, 1992, pp.109-118, ISSN 0167-6393, https://doi.org/10.1016/0167-6393(92)90005-R.
[17] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," Computer Speech & Language, vol.26, issue 1, 2012, pp.20-34, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2011.03.003.
[18] Y. -R. Chien, D. D. Mehta, J. Guðnason, M. Zañartu and T. F. Quatieri, "Evaluation of Glottal Inverse Filtering Algorithms Using a Physiologically Based Articulatory Speech Synthesizer," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 8, pp. 1718-1730, Aug. 2017, doi: 10.1109/TASLP.2017.2714839.
[19] A. Sasou and K. Tanaka, "Glottal excitation modeling using HMM with application to robust analysis of speech signal", In ICSLP-2000, vol.4, 704-707.
[20] A. Sasou, M. Goto, S. Hayamizu, K. Tanka, "An auto-regressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition," Proc. of ICASSP I, pp.237-240, 2005.
[21] A. Sasou, "Automatic Topology Generation of Glottal Source HMM," Proc. of Interspeech 2012.
[22] A. Mohamed et al., "Self-Supervised Speech Representation Learning: A Review," in IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1179-1210, Oct. 2022, doi: 10.1109/JSTSP.2022.3207050.
[23] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol.349, no.6245, pp. 255-260, 2015.
[24] S.-w. Yang, et al., "SUPERB: Speech Processing Universal PERformance Benchmark," Proc. Interspeech 2021, pp.1194-1198, doi: 10.21437/Interspeech.2021-1775.
[25] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," in Proc. ICASSP'92, vol. 1, pp.573.576, 1992.
[26] T.Jitsuhiro, T.Matsui, S.Nakamura, "Automatic Generation of Non-Uniform Context-Dependent HMM Topologies Based on The MDL Criterion," in Proc. of EUROSPEECH2003, pp.2721-2724,2003.
[27] M. Putzer and W. Barry, "Saarbrucken Voice Database", Institute of Phonetics, Univ. of Saarland, Accessed March 2021 from http://www.stimmdatenbank.coli.uni-saarland.de/
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, "Attention Is All You Need," Proc. of NIPS, 2017.
[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. of ICLR, 2015.
[30] M. Riviere, A. Joulin, P.-E. Mazare, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in ICASSP, 2020, pp. 7414–7418.
[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in NeurIPS, 2020.
[32] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," arXiv preprint arXiv:2106.07447, 2021.
[33] I. Zenkov, "Parallel is All You Want: Combining Spatial and Temporal Feature Represention of Speech Emotion by Parallelizing CNNs and Transformer-Encoders," Github, https://github.com/IliaZenkov/transformer-cnn-emotion-recognition, 2020.
[34] A. Sasou, "Voice-pathology analysis based on AR-HMM," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea (South), 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820679.