The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection

Saska Tirronen, Sudarsana Reddy Kadiri, and Paavo Alku, *Finland*

Summary: Automatic voice pathology detection is a research topic, which has gained increasing interest recently. Although methods based on deep learning are becoming popular, the classical pipeline systems based on a two-stage architecture consisting of a feature extraction stage and a classifier stage are still widely used. In these classical detection systems, frame-wise computation of mel-frequency cepstral coefficients (MFCCs) is the most popular feature extraction method. However, no systematic study has been conducted to investigate the effect of the MFCC frame length on automatic voice pathology detection. In this work, we studied the effect of the MFCC frame length in voice pathology detection using three disorders (hyperkinetic dysphonia, hypokinetic dysphonia and reflux laryngitis) from the Saarbrücken Voice Disorders (SVD) database. The detection performance was compared between speaker-dependent and speaker-independent scenarios as well as between speaking task -dependent and speaking task -independent scenarios. The Support Vector Machine, which is the most widely used classifier in the study area, was used as the classifier. The results show that the detection accuracy depended on the MFFC frame length in all the scenarios studied. The best detection accuracy was obtained by using a MFFC frame length of 500 ms with a shift of 5 ms.

Key Words: Voice pathology−Pathology detection−Speech analysis−MFCC−SVM.

INTRODUCTION

In addition to its linguistic content, speech contains vast amounts of other information about, for example, the speaker's vocal emotions and state of health. This so called paralinguistic information can be extracted by automatic systems from the acoustic speech signal and utilized, for example, in medical diagnosis. The current study investigates the utilization of paralinguistic information by focusing on the automatic detection of voice pathologies from speech signals.

Different voice pathology detection systems have been developed based on two main approaches: traditional pipeline systems and end-to-end systems. Traditional pipeline systems use a two-stage structure consisting of a feature extraction stage and a classification stage. In the feature extraction stage, acoustic features are computed to express the input voice signal in a parametric form. Many different acoustic features have been investigated in the study area including, for example, mel-frequency cepstral coefficients (MFCCs),[1−3] linear predictive cepstral coefficients (LPCCs)[4−6] and perceptual linear prediction (PLP)[7,8] coefficients. Acoustical features are typically computed frame wise and these features are processed with statistical functionals (eg, mean, standard deviation, maximum) to express the corresponding features' values over a longer spoken unit (eg, phone, syllable, word). In the second stage, the features are used by a machine learning (ML) classifier to classify the voice signal either as healthy or pathological. Examples of classifiers used in the study area are support vector machine (SVM), K-nearest neighbors (KNN), Gaussian mixture models (GMM), and linear discriminant analysis (LDA).[9−13] In end-to-end systems, the use of handcrafted features is replaced by training deep learning models to directly map the raw speech signal or the spectrogram to the output labels (healthy vs. pathological). Although end-to-end systems are becoming increasingly popular and they have been used, for example, in the detection of Parkinson's disease,[14−17] they suffer from their need for large amounts of data in model training. Therefore, the traditional pipeline system is still an effective approach in the detection of voice pathologies and will be used also in this study. The focus of the current study is in the feature extraction stage.

Mel-frequency cepstral coefficients (MFCCs) are the most widely used acoustic features in all areas of speech technology and they have also been used widely in automatic voice pathology detection with traditional pipeline systems.[1−3,9,18,19] In the study area of automatic voice pathology detection, MFCCs have not only been used as an individual feature set but they are also included in widely used large feature sets such as openSMILE,[20] articulation features[14,15] and glottal source features.[21−24] In the extraction of MFCCs, framing is performed for a pre-emphasized audio signal by splitting the signal into segments of equal length by using a window function. For each of the framed signals, the logarithmic power spectrum is computed by applying the discrete Fourier transform and taking its logarithm. MFCCs are then obtained for each frame by conducting mel-scaled filter bank analysis, and by applying discrete cosine transform (DCT).

The focus in many previous investigations in the detection of voice pathologies (eg,[25,26]) has been in new features. In these studies, MFCCs have been almost exclusively used as the default reference features. Although MFCCs have been

Accepted for publication March 21, 2022.
From the Department of Signal Processing and Acoustics, Aalto University, Finland.
Address correspondence and reprint requests to Sudarsana Reddy Kadiri. E-mail: sudarsana.kadiri@aalto.fi
Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2022.03.021

used widely in the study area of voice pathology detection, it is surprising that the effect of one basic attribute, the frame length in the MFCC computation, has not been studied systematically. Therefore, the main focus of this study is to investigate how the detection of voice pathologies is affected when the MFCC feature extraction is computed using different frame lengths while keeping the shift between the frames at a default constant small value of 5 ms[3,27] and by using the mean as a statistical functional to combine frame-wise MFCCs into an utterance-wise feature vector. Intuitively, this corresponds to using varying levels of temporal smoothing in the frame-wise MFCC-based feature extraction phase of the detection system. The detection problem studied is a binary classification task in which healthy voice is distinguished from pathological voice. Pathological voice is represented by three voice disorders (hyperkinetic dysphonia, hypokinetic dysphonia, and reflux laryngitis). These three disorders were selected because they are common voice pathologies and because there are open databases (eg,[28]), which provide sufficient amounts of training data of both female and male speakers to train ML models to detect these three disorders. As the classifier, we use the support vector machine (SVM), which is the most widely used classifier in the study area of voice pathology detection.[8,29−31] The main highlights of this study are as follows:

- A systematic investigation is conducted to study the effect of the MFCC frame length on the automatic voice pathology detection.
- The effect of the MFCC frame length on pathology accuracy is compared between a gender-dependent and a gender-independent scenario and between a speaking task -dependent and a speaking task -independent scenario.

The paper is organised as follows. Section Database describes the SVD database and its speaking tasks and pathologies used. The methodology of the present study is described in Section Methodology, which includes the feature extraction, the details of the classifier, the experiments carried out, and the evaluation metrics considered. The results of the experiments are reported in Section Results. Finally, Section Discussion and conclusions summarizes the study.

DATABASE

The SVD database was recorded at the Institut für Phonetik at Saarland University and the Phoniatry Section of the Caritas Clinic St. Theresia in Saarbrücken.[28,32] The total number of samples in the database is 2225, of which 869 are healthy and 1356 are pathological. There are altogether 71 pathologies in the database. The database contains recordings from male and female speakers in several speaking tasks.

The pathologies that were examined in this study are hyperkinetic dysphonia, hypokinetic dysphonia and reflux laryngitis. We used recordings from both male and female speakers, and from the following two speaking tasks: (1) pronunciation of a German sentence ”Guten Morgen, wie geht es Ihnen” (”Good morning, how are you?”), and (2) sustained pronunciation of three vowels (/a/, /i/, /u/) in a constant ’neutral’ pitch.

The number of speech samples studied in this paper is shown in Figure 1 for the sentence production task. Figure 2 shows the number of speech samples for the production of the vowel /a/ (for the other two vowels, the histogram is roughly the same). As there are more recordings from healthy speakers than from patients in the selected three pathologies, we balanced the class sizes by randomly selecting a correctly sized subset of healthy samples.

METHODOLOGY

Figure 4 shows the flow diagram of the traditional pipeline system used in this study in the pathology detection. In the following two sub-sections, the two stages of the system are first described after which we describe the experiments that

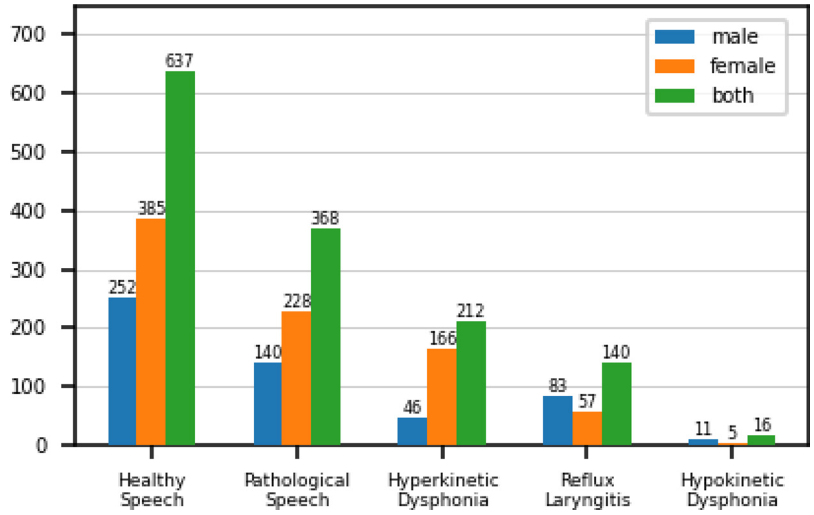

FIGURE 1. Number of speech samples in the sentence pronunciation task by male and female speakers of healthy and pathological speech.

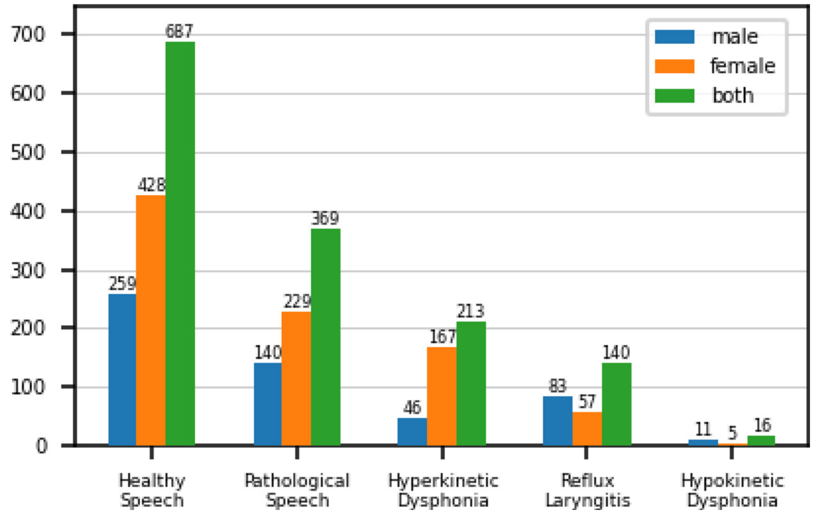

FIGURE 2. Number of speech samples in the vowel pronunciation task (for the vowel /a/) by male and female speakers of healthy and pathological speech.

we designed to study the effect of the MFCC frame length using the detection system.

Feature Extraction

As the pre-processing step, audio files were first down-sampled from the original sampling frequency of 50 kHz to 16 kHz. Silence parts were next detected and removed by using a volume threshold of 1% and by requiring the silence to last for at least 0.1 seconds in order to be removed.

MFCCs were extracted from each input speech signal in frames by using five values (25 ms, 50 ms, 100 ms, 200 ms and 500 ms) for the frame length. The frame shift was kept at a constant value of 5 ms for all the five frame lengths and the Hamming window was applied to the frames. The first 13 cepstral coefficients (including the 0^{th} coefficient) were computed in each frame. Finally, the utterance-based MFCC vector was obtained by computing the mean from the frame-wise MFCCs. Note that the computation of MFCCs described above corresponds to using only static MFCCs. This choice was made in order to understand the effect of the frame length in its purest form, that is, without combining information across neighboring frames as used in dynamic features (ie, first and second derivatives). The MFCC extraction was implemented using the Bob toolkit.[33,34]

Frame lengths longer than 500 ms could not be used because of the short durations of some of the vowels in the SVD database. The histogram of the signal lengths for the vowels and sentences is shown in Figure 3.

Classifier

The voice pathology detection was carried out using the SVM classifier, which was implemented using the Scikit-learn library.[35] We used the radial basis function (RBF) kernel $\kappa(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{1}{D\cdot Var(X)} \parallel \mathbf{u}-\mathbf{v} \parallel^{2}\right)$, where D is the dimensionality of the feature space, and $Var(X)$ represents

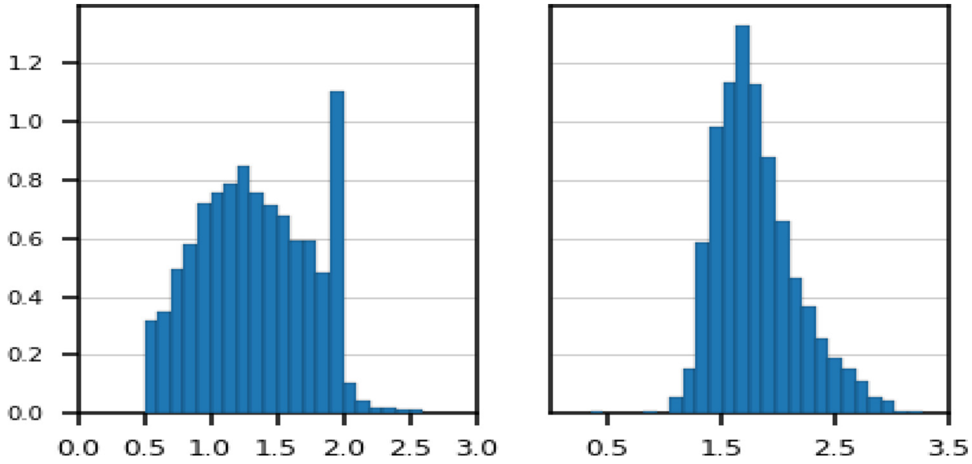

FIGURE 3. Histograms for the signal length in the vowel production task (left panel) and in the sentence production task (right panel). The x-axis corresponds to time in seconds. The y-axis is normalized so that the histogram sums to 1.

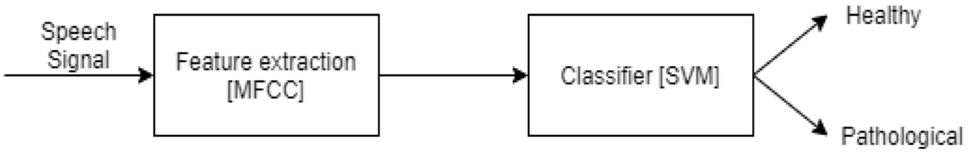

FIGURE 4. The voice pathology detection system. The system consists of two main stages: feature extraction and classification. SVM refers to Support Vector Machine.

the variance of the training data. The generalization parameter C was set to 1.0.

All experiments were carried out using 20-fold cross-validation, where one of the folds was used for testing, while the others were used for training. The mean and standard deviation of the MFCC training data were used to perform Z-score normalization for both the training and testing data.

Experiments

In order to study the effect of the MFCC frame length on pathology detection, three different experiments were conducted and these experiments will be described in this section. The experiments were designed to understand the effect of two attributes (speaker gender and speaking task) on the detection performance when the length of the MFCC frame was varied. We selected gender and speaking task as attributes of interest because they are the most common general factors which are typically varied in collecting speech to voice pathology databases (such as SVD) and which are available in the databases as labels to separate the speech data into sub-groups of different acoustical properties.

A. The gender and speaking task -independent scenario

The first scenario corresponds to a general experiment type, which is typically used in the study area[3,18,27]: we studied the effect of the frame length in a general scenario where we used the data of both genders and of all four speaking tasks available in the SVD database to train and test the pathology detection system.

B. The speaking task -dependent scenario

In the second experiment, we examined if the effect of the frame length in the pathology detection is affected by the speaking task. The data was divided into four subsets (three vowel production tasks, one sentence production task) according to the speaking task used in the recordings. Each subset contained recordings of one speaking task only. The classifier was trained and evaluated for each of these subsets individually. Both genders were included in the data.

C. The gender-dependent scenario

For the third experiment, the data was divided into two subsets: one containing recordings from male subjects and the other one containing recordings from female subjects. As the amount of data from the female subjects is considerably larger than from the male subjects, we balanced both data sets to have an equal number of recordings for each pathology and speaking task.

Evaluation

Classifier performance was evaluated by using the following six metrics[8,27]: accuracy, precision (PREC), recall (REC), area under the receiver operating characteristic curve (AUC), F1-score, and equal error rate (EER). The better the classifier's performance the higher the values of the first five metrics will be. For the last metric (EER), a lower value corresponds to a better performing system.[8,27]

RESULTS

In this section, the results of each experiment are described and discussed. The obtained performance metrics results are given in Tables 1-3. We report metrics in these tables as in review article.[27] Only the classification accuracy is visualized, as the other performance metrics showed similar trends.

The gender and speaking task -independent scenario

The results obtained in the gender and speaking task -independent scenario are shown in Figure 5. It can be observed

TABLE 1.
Performance Metric Values For the Gender and Speaking Task -Independent Scenario

Frame [ms]	Accuracy [%]	PREC	REC	AUC	F1	EER
			Gender and speaking task -independent scenario			
25	60.3 ± 4.6	0.60	0.60	0.65	0.60	0.41
50	63.4 ± 3.9	0.63	0.63	0.68	0.63	0.36
100	63.8 ± 2.6	0.64	0.64	0.69	0.64	0.36
200	63.9 ± 4.5	0.64	0.64	0.69	0.64	0.36
500	66.4 ± 3.3	0.67	0.66	0.72	0.66	0.34

PREC refers to precision, REC refers to recall, AUC refers to area under the receiver operating characteristic curve, F1 refers to F1-score, and EER refers to equal error rate.

TABLE 2.
Performance Metric Values For the Speaking Task-Dependent Scenario

Frame [ms]	Accuracy [%]	PREC	REC	AUC	F1	EER
Speaking task-dependent scenario: ′/a/′						
25	62.7 ± 7.2	0.63	0.63	0.68	0.62	0.36
50	65.0 ± 6.3	0.65	0.65	0.70	0.65	0.36
100	65.4 ± 6.9	0.66	0.65	0.70	0.65	0.35
200	66.7 ± 6.6	0.67	0.67	0.73	0.66	0.34
500	68.5 ± 9.6	0.69	0.68	0.75	0.68	0.33
Speaking task-dependent scenario:′/i/′						
25	58.6 ± 6.4	0.59	0.59	0.63	0.58	0.41
50	59.6 ± 5.9	0.60	0.60	0.64	0.59	0.38
100	64.0 ± 8.3	0.64	0.64	0.69	0.64	0.35
200	64.1 ± 7.9	0.64	0.64	0.68	0.64	0.37
500	65.2 ± 7.1	0.66	0.65	0.69	0.65	0.35
Speaking task-dependent scenario:′/u/′						
25	56.2 ± 10.7	0.56	0.56	0.60	0.56	0.42
50	59.1 ± 9.7	0.59	0.59	0.62	0.59	0.40
100	54.9 ± 8.2	0.55	0.55	0.60	0.55	0.42
200	60.6 ± 6.7	0.61	0.61	0.64	0.60	0.40
500	60.8 ± 7.6	0.61	0.61	0.66	0.60	0.37
Speaking task-dependent scenario: ′sentence′						
25	67.2 ± 8.2	0.68	0.67	0.72	0.67	0.35
50	67.5 ± 6.5	0.68	0.68	0.74	0.67	0.32
100	71.3 ± 7.3	0.72	0.71	0.77	0.71	0.28
200	68.6 ± 7.7	0.69	0.69	0.74	0.68	0.32
500	75.1 ± 9.3	0.75	0.75	0.80	0.75	0.27

PREC refers to precision, REC refers to recall, AUC refers to area under the receiver operating characteristic curve, F1 refers to F1-score, and EER refers to equal error rate.

from this figure that using the MFCC frame length of 500 ms yielded a clearly higher classification accuracy than using a frame length of 20 ms. The rising trend in accuracy is consistent also in between 20 ms and 500 ms. As shown in Table 1, the mean accuracy values in the gender and speaking task -independent system are 60.3%, 63.4%, 63.8%, 63.9% and 66.4% for the frame lengths of 25 ms, 50 ms, 100 ms, 200 ms and 500 ms, respectively.

The speaking task -dependent scenario

The accuracy is shown as a function of the MFCC frame length in Figure 6 for the speaking task -dependent scenario. As can be observed, the classification accuracy was again clearly higher for the frame length of 500 ms compared to the frame length of 25 ms. The largest absolute difference of 7.9 percentage points was observed for the sentence production task, for which the mean accuracy was 67.2%, 67.5%,

TABLE 3.
Performance Metric Values For the Gender-Dependent Scenario

Frame [ms]	Accuracy [%]	PREC	REC	AUC	F1	EER
Gender-dependent scenario: ′male′						
25	60.6 ± 7.6	0.61	0.61	0.63	0.60	0.40
50	59.5 ± 6.5	0.60	0.60	0.65	0.59	0.40
100	63.4 ± 7.2	0.64	0.63	0.68	0.63	0.36
200	61.2 ± 5.0	0.61	0.61	0.68	0.61	0.38
500	63.5 ± 5.9	0.64	0.64	0.69	0.63	0.36
Gender-dependent scenario: ′female′						
25	57.3 ± 7.1	0.57	0.57	0.61	0.57	0.42
50	64.1 ± 5.2	0.64	0.64	0.69	0.64	0.36
100	65.3 ± 6.2	0.66	0.65	0.72	0.65	0.34
200	66.4 ± 4.7	0.67	0.66	0.72	0.66	0.35
500	66.5 ± 6.2	0.67	0.66	0.72	0.66	0.33

PREC refers to precision, REC refers to recall, AUC refers to area under the receiver operating characteristic curve, F1 refers to F1-score, and EER refers to equal error rate.

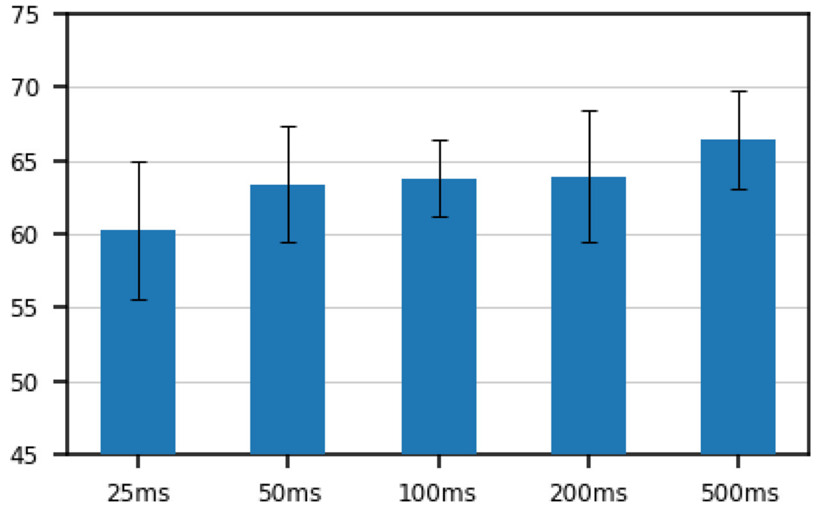

FIGURE 5. Detection accuracy (in %) as a function of the frame length in the gender and speaking task -independent scenario. Bar height shows the mean and tail shows the standard deviation of the accuracies over 20-fold cross-validation. Both genders and all speaking tasks were included in the data.

71.3%, 68.6% and 75.1% for the frame lengths of 25 ms, 50 ms, 100 ms, 200 ms and 500 ms, respectively (see Table 2). In addition, Figure 6 shows that for all the frame lengths studied, the performance of the detection varied between the three vowels and that accuracy was always lowest in /u/, second lowest in /i/ and highest in /a/.

The gender-dependent scenario

Figure 7 shows the detection accuracy as a function of the MFCC frame length for the gender-dependent system. It can be seen from this figure that the accuracy is clearly higher for the frame length of 500 ms compared to the frame length of 25 ms. In addition, this improvement in accuracy is observed for both genders and particularly for the female speakers. The system trained/tested using male speech seems to perform better than the one trained/tested using female speech for the shortest MFCC frame length of 25 ms. However, when the length of the MFCC frame increases the accuracy of the detection system trained/tested using female speech improves and becomes better than that of the system based on male speech. The mean accuracy values of the system trained/tested using female speech were 57.3%, 64.1%, 65.3%, 66.4% and 66.5% for the frame lengths of 25 ms, 50 ms, 100 ms, 200 ms and 500 ms, respectively (see Table 3). The absolute difference between the two extreme frame lengths is 9.2 percentage points.

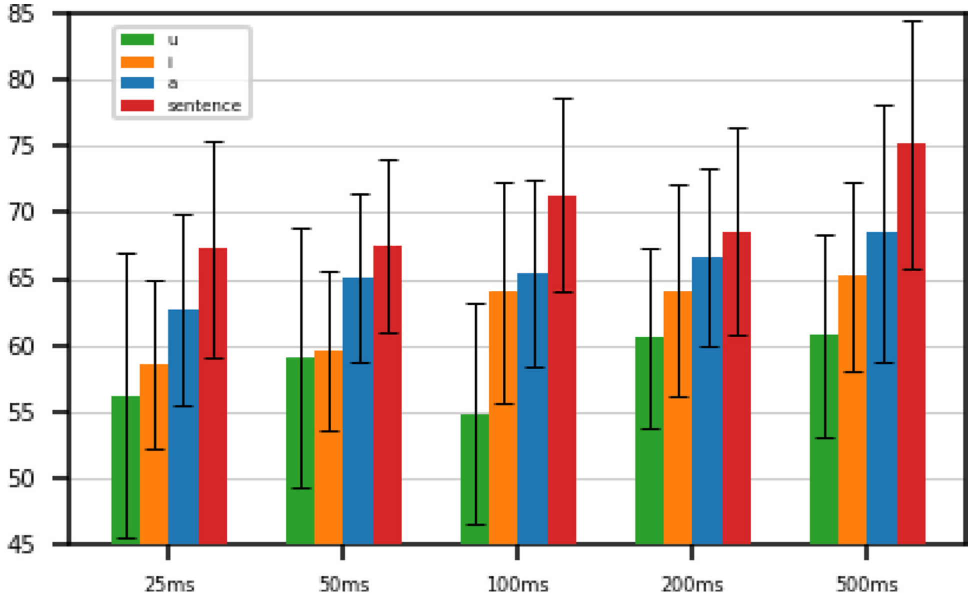

FIGURE 6. Detection accuracy as a function of the frame length in the speaking task -dependent scenario. Bar height shows the mean and tail shows the standard deviation of the accuracies over 20-fold cross-validation. Data from both genders was included.

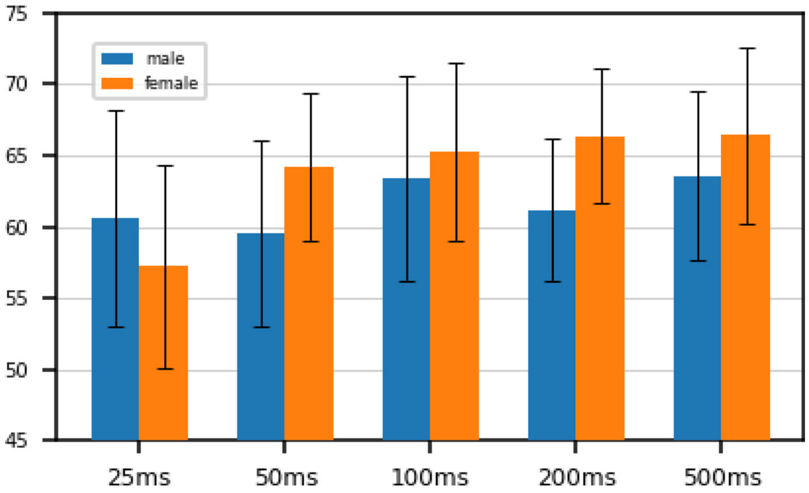

FIGURE 7. Detection accuracy as a function of the frame length in the gender-dependent scenario. Bar height shows the mean and tail shows the standard deviation of the accuracies over 20-fold cross-validation. All speaking tasks were included in the data.

DISCUSSION AND CONCLUSIONS

The MFCC features are widely used in the detection of pathological speech when the classical pipeline approach is used as the system architecture in the binary classification task. Despite the fact that MFCCs have been used widely as speech features (or they have been used as default reference features in studies proposing new features), the effect of the frame length in the MFCC computation has not been systematically studied before in the detection of speech pathologies. Therefore, the goal of the present study was to investigate the effect of the MFCC frame length in the pathology detection using three speech pathologies (hyperkinetic dysphonia, hypokinetic dysphonia, and reflux laryngitis) from the SVD database.

Three experiments were carried out to test the effect of the frame length in three different scenarios: the gender and speaking task -independent scenario, the speaking task -dependent scenario and the gender-dependent scenario. The results showed that by increasing the MFCC frame length from its default value of 25 ms to a much lager value of 500 ms yielded improved classification accuracy in all the scenarios. The largest effect in terms of the accuracy gain in percentage points was observed in the gender-dependent scenario for the system, which was trained and tested using speech of female talkers. In this case, the difference in the detection accuracy between the two frame lengths of 25 ms and 500 ms was as large as 9.2 percentage points. In the speaking task -dependent scenario, the detection accuracy was higher for the sentence recordings than for any of the vowel recordings. In the gender-dependent scenario, the results indicated that the detection system which was trained and tested using male speech performed better than the one which was trained and tested using female speech when the shortest MFCC frame length of 25 ms was used. However, the gender-dependent detection system based on female speech improved in accuracy when the MFCC frame length was increased and was better than the system based on male speech for the largest frame length of 500 ms.

It is to be noted that MFCCs were developed many years ago as a feature extraction method for automatic speech recognition (ASR)[36] and they have since been used widely in many other areas of speech and audio technology. Using a short frame length of a few tens of milliseconds (eg 25 ms) is justified in ASR because a frame length of this size is long enough in order to get enough data samples to model vocal tract information with adequate spectral resolution but short enough in order not to span multiple phones. In the case of the current study, however, the default length of 25 ms gave the poorest detection performance and the use of a much longer frame length (of 500 ms) in the MFCC computation was best. The improved performance given by longer frame lengths in the detection problem of the current study can be explained by the nature of the three pathologies studied: All three voice disorders investigated are laryngeal and therefore accurate capturing of vocal tract information for each phone (which is known to be the advantage of the MFCCs) is not as necessary in the current detection task as it would be, for example, in ASR. In fact, it can be argued that particularly in the detection of sentences (which consist of multiple phones), the tendency of the default MFCC feature extraction to capture vocal tract information of each phone of the utterance may deteriorate the automatic pathology detection if the underlying pathology is vocal fold -related as in the current study. In conclusion, the current results show that the detection of laryngeal voice pathologies was improved by lengthening the MFCC frame from its default duration of 25 ms to 500 ms because this lengthening of the MFCC frame reduced the biasing of the features by individual phones and gave more data samples for the frame-wise computation of MFCCs.

Acknowledgments

This study was funded by the Academy of Finland (project no. 330139). Aalto ScienceIT provided the computational resources.

REFERENCES

1. Watts CR, Awan SN. Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *J Speech Lang Hearing Res*. 2011;54:1525–1537.
2. Godino-Llorente JI, Vilda PG. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Trans Biomed Eng*. 2004;51:380–384.
3. García JAG, Moro-Velázquez L, Godino-Llorente JI. On the design of automatic voice condition analysis systems. part I: review of concepts and an insight to the state of the art. *Biomed Signal Process Control*. 2019;51:181–199.
4. Godino-Llorente JI, Aguilera-Navarro S, Vilda PG. LPC, LPCC and MFCC parameterisation applied to the detection of voice impairments. *Proc Interspeech*. 2000:965–968.
5. Saldanha JC, Ananthakrishna T, Pinto R. Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features. *J Med Imaging Health Informat*. 2014;4:168–173.
6. Parsa V, Jamieson DG. Identification of pathological voices using glottal noise measures. *J Speech Lang Hearing Res*. 2000;43:469–485.
7. Little MA, Costello DA, Harries ML. Objective dysphonia quantification in vocal fold paralysis: Comparing nonlinear with classical measures. *J Voice*. 2011;25:21–31.
8. Kadiri SR, Alku P. Analysis and detection of pathological voice using glottal source features. *IEEE J Sel Top Signal Process*. 2020;14:367–379.
9. Godino-Llorente JI, Vilda PG, Blanco-Velasco M. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. *IEEE Trans Biomed Eng*. 2006;53:1943–1953.
10. Arjmandi MK, Pooyan M, Mikaili M, et al. Identification of voice disorders using long-time features and support vector machine with different feature reduction methods. *J Voice*. 2011;25:e275–e289.
11. Akbari A, Arjmandi MK. An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features. *Biomed Signal Process Control*. 2014;10:209–223.
12. Alhussein M, Muhammad G. Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access*. 2018;6:41034–41041.
13. Hegde S, Shetty S, Rai S, et al. A survey on machine learning approaches for automatic detection of voice disorders. *J Voice*. 2018.
14. Vásquez-Correa JC, Orozco-Arroyave JR, Noth E. Convolutional neural network to model articulation impairments in patients with Parkinson's disease. *Proc Interspeech*. 2017:314–318.
15. Vásquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, et al. A multitask learning approach to assess the dysarthria severity in patients with Parkinson's disease. In: *Proc Interspeech*20172017:314–318.
16. Vásquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, et al. Multimodal assessment of Parkinson's disease: A deep learning approach. *IEEE J Biomed Health Informat*. 2019;23:1618–1630.
17. Arias-Vergara T, Vásquez-Correa JC, Orozco-Arroyave JR, et al. Unobtrusive monitoring of speech impairments of Parkinson's disease patients through mobile devices. In: *Proc IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 20182018:6004–6008.
18. Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, et al. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Trans Biomed Eng*. 2011;58:370–379.
19. Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust Speech Signal Proces*. 1980;28:357–366.
20. Eyben F, Wöllmer M, Schuller B. Opensmile: the Munich versatile and fast open-source audio feature extractor. In: *Proc ACM International Conference on Multimedia*20102010:1459–1462.
21. Narendra NP, Alku P. Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. *Comput Speech Lang*. 2021;65:101117.
22. Kadiri SR, Alku P, Yegnanarayana B. Analysis and classification of phonation types in speech and singing voice. *Speech Commun*. 2020;118:33–47.
23. Kadiri SR, Alku P. Glottal features for classification of phonation type from speech and neck surface accelerometer signals. *Comput Speech Lang*. 2021;70:101232.
24. Kadiri SR, Alku P. Mel-frequency cepstral coefficients of voice source waveforms for classification of phonation types in speech. In: *Proc. Interspeech 2019*20192019:2508–2512.
25. Arias-Londoño JD, Godino-Llorente JI, Markaki M, et al. On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices. *Logopedics Phoniatrics Vocol*. 2011;36:60–69.
26. Vilda PG, Fernández-Baíllo R, Biarge MVR, et al. Glottal source biometrical signature for voice pathology detection. *Speech Commun*. 2009;51:759–781.
27. García JAG, Moro-Velázquez L, Godino-Llorente JI. On the design of automatic voice condition analysis systems. part II: review of speaker recognition techniques and study on the effects of different variability factors. *Biomed Signal Process Control*. 2019;48:128–143.
28. Pützer M., Barry W.J. Saarbrücken voice database, institute of phonetics, univ. of saarland. 2010. http://www.stimmdatenbank.coli.uni-saarland.de/(Last viewed April 1, 2021).
29. Gelzinis A, Verikas A, Bacauskiene M. Automated speech analysis applied to laryngeal disease categorization. *Comput Methods Programs Biomed*. 2008;91:36–47.
30. Fonseca ES, Guido RC, Scalassara PR, et al. Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders. *Comput Biol Med*. 2007;37:571–578.
31. Akbari A, Arjmandi MK. Employing linear prediction residual signal of wavelet sub-bands in automatic detection of laryngeal pathology. *Biomed Signal Process Control*. 2015;18:293–302.
32. Pützer M, Barry WJ. Instrumental dimensioning of normal and pathological phonation using acoustic measurements. *Clin Linguist Phonet*. 2008;22:407–420.
33. Anjos A, Günther M, de Freitas Pereira T, et al. Continuously reproducing toolchains in pattern recognition and machine learning experiments. *International Conference on Machine Learning (ICML)* 20172017:1–8.
34. Anjos A., El-Shafey L., Wallace R., Günther M., et al. Bob: a free signal processing and machine learning toolbox for researchers. Proceedings of the 20th ACM International Conference on Multimedia 2012:1449−1452.
35. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Research*. 2011;12:2825–2830.
36. Povey D, Ghoshal A, Boulianne G, et al. The Kaldi Speech recognition toolkit. *Proc ASRU*. 2011:1–4.