Cross-lingual dysphonic speech detection using pre-trained speaker embeddings

Aziz Dosti Ali Hama Salih, Dávid Sztahó

Budapest University of Technology and Economics,
Magyar tudósok körútja 2., 1117 Budapest, Hungary
azizd@edu.bme.hu, sztaho.david@vik.bme.hu

Abstract: In this study, cross-lingual binary classification and severity estimation of dysphonic speech have been carried out. Hand-crafted acoustic feature extraction is replaced by the speaker embedding techniques used in the speaker verification. Two state of art deep learning methods for speaker verification have been used: the X-vector and ECAPA-TDNN. Embeddings are extracted from speech samples in Hungarian and Dutch languages and used to train Support Vector Machine (SVM) and Support Vector Regressor (SVR) for binary classification and severity estimation, in a cross-language manner. Our results were competitive with manual feature engineering, when the models were trained on Hungarian samples and evaluated on Dutch samples in the binary classification of dysphonic speech and outperformed in estimating the severity level of dysphonic speech. Moreover, our model achieved 0.769 and 0.771 in Spearman and Pearson correlations. Also, our results in both classification and regression were superior compared to manual feature extraction technique when models were trained on Dutch samples and evaluated on Hungarian samples with only a limited number of samples are available for training. An accuracy of 86.8% was reached with features extracted from embedding methods, while the maximum accuracy using hand-crafted acoustic features was 66.8%. Overall results show that Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) performs better than the former X-vector in both tasks.

Keywords: dysphonic, cross-lingual, speaker verification, embedding, X-vector, ECAPA-TDNN

1 Introduction

Biomarkers are implications that indicate a medical condition observed in the patient. They contain a wide range of medical marks that can be accurately measured. They refer to any sign or indication which may lead to predicting a disease, starting from a basic blood test, and going to more complicated tests performed in a specialized laboratory (Strimbu & Tavel, 2010).

Even though the human voice is the main source for communication and social interactions among individuals, it also carries further information about the identity and health status of a person (Grossmann et al., 2013). Speech is considered as a biomarker due to its ability to make health specialists diagnose diseases based on a patient's

speech. Many diseases affect the speech-production organs in the human body, making them face difficulties in producing normal speech (Lin et al., 2020). Dysphonia, also called hoarseness, is a condition in which the produced voice has alterations in the quality, pitch and loudness. It affects specific groups of people, mostly elderly ones, teachers and individuals with extensive use of vocal voices. Nearly one out of three people at some time in their life will be diagnosed as having dysphonia (Schwartz et al., 2009; Stemple et al., 2018; Van Houtte et al., 2011). A person diagnosed with dysphonia requires regular consultations with clinicians to assess the severity of their condition. The implementation of these measures is expected to result in significant costs for both work environments and healthcare systems due to the disruption caused by the illness and the need for medical care for affected individuals. (Cohen et al., 2012; Schwartz et al., 2009).

Therefore, using artificial intelligence to analyse human speech to detect disorders has been extensively researched by academia and clinicians. Voice samples from both healthy and disordered persons were converted to feature vector representations using a digital signal processing technique (Hegde et al., 2019).

Various research has been conducted using extracted features from speech samples to train various machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest, Decision Tree and many others to distinguish between healthy and dysphonic speech (Hadjaidji et al., 2021; Syed et al., 2021; Wu et al., 2017).

The vast development of deep learning techniques during the last few decades led to advancement in many research areas including computer vision, natural language processing and speech and speaker recognition. Speaker embedding methods originally used for speaker verification and identification can capture speaker-related characteristics which are used to identify speakers based on their speech signals (Bimbot et al., 2004; Togneri & Pullella, 2011).

Some researchers adopted speaker embedding techniques for classifying disordered speech from a normal one. The objective of these methods is that they don't require hand-crafted feature extraction and engineering (Egas-López et al., 2022; Scheuerer et al., 2021).

In this research, we adapted two state of art speaker embedding techniques, the X-vector and Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) to extract features from speech samples in different languages and use them for classifying dysphonia. SVM and SVR algorithms are trained on features extracted from samples from a Hungarian dataset and tested on features extracted from a Dutch dataset and vice versa. The result shows that the embedding techniques are suitable for cross-language detection of dysphonic speech.

In section 2, we will present related works using feature engineering and speaker embedding techniques. A detailed description of the used dataset and methods will be examined in section three. Section 4 will illustrate the obtained results from classification and regression. Lastly, the discussion and conclusion will be included in 5 and 6 sections, respectively.

2 Previous related work

Machine learning models have shown great advancements in many medical areas such as cancer and tumour detection. Analysing speech to diagnose disease has been an active research area; there have been enormous machine learning approaches to classify dysphonic speech from a healthy control.
Al-Dhief et al. extract features from Mel-Frequency Cepstral Coefficients (MFCC) in speech samples that were found to be effective in pathological voice analysis. Features extracted using vowel /a/ from the German dataset Saarbrucken Voice Database (SVD). They achieved an accuracy of 85% using Online Sequential Extreme Learning Machine (OSELM) algorithm with 200 hidden units, an improved version of the ELM algorithm (AL-Dhief et al., 2020). Their work (Dankovičová et al., 2018) presents results of identifying dysphonic samples from healthy control speech using Random Forest, SVM and K-nearest neighbours classifiers. A total of 1560 features were extracted from the vowels /a/, /i/ and /u/, 520 features for each vowel, and dimensionality reduction of features was performed using Principal Component Analysis (PCA). The highest achieved accuracy was by Support Vector Machine (SVM) classifier on the male samples, and the obtained accuracy was 91%. In another study (Awan & Roy, 2006), predicting the severity of dysphonic speech was researched in which, time and spectral-based features derived from sustained vowels were considered with stepwise multiple regression. The reported results were 0.880 and 0.775 for mean R and mean R2, respectively. The study shows that the four-variable model included time and spectral-based acoustic measures were able to strongly predict perceived severity.

Several studies have been conducted by adopting deep learning methods for speaker embedding for identifying disordered speech. These approaches do not require handcrafted feature extraction, as they are able to include speaker characteristic features in the embedding. In (Scheuerer et al., 2021) implementation of two models of X-vectors based on Mel-frequency cepstral coefficient (MFCC) and gammatone frequency cepstral coefficients (GFCC) carried out for both regression and binary classification on SVD dataset. According to their findings, the GFCC-based multi-layer perceptron was the best, reaching 0.8810 and 0.8810 ROG AUC scores for both regression and classification, respectively. Pre-trained i-vector and X-vector are used for classifying Parkinson's disease and obstructive sleep apnnea, and they perform better than handcrafted feature extractions. Moreover, X-vector performs better when there is a domain mismatch between the train and test speech samples (Botelho et al., 2020).

Relating to cross-lingual voice disorder detection, some research has been carried out. Cross-lingual detection of voice disorders was implemented using speech samples from Spanish, Czech and Dutch. Different kinds of training and testing cases were performed using these languages. According to their experiments, the highest accuracies achieved in the test set were nearly 70%, and 60% on Czech and German samples, respectively, when the algorithm was trained on Spanish samples. They also examined the improvement of accuracies by moving speech samples from the target database to the training set. They reported nearly 30% improvement in German samples by adding only 20% of German samples to the training set (Orozco-Arroyave et al., 2016).

In (Shinohara et al., 2017), monolingual evaluation of the pitch-related features has been performed in German, Spanish and Czech languages using normal speech utterances to identify voice disorders.

Results achieved in this study are compared to previous work that has been done by (Sztahó et al., 2022), performing a cross-lingual evaluation of classifying and estimating the severity level of dysphonic speech in Hungarian and Dutch languages using hand-crafted acoustic features. The results in that study show training the machine learning algorithms on features extracted from Hungarian utterances and testing on Dutch samples is possible. The accuracy of 88% in test samples was achieved with acoustic features extracted from entire utterances with phoneme level features of \E\. Results of severity estimation were 0.72 and 0.79 for Pearson correlation and RMSE, respectively.

3 Methods

3.1 Description of Databases

Hungarian and Dutch samples from two dysphonic speech datasets were used for the experiment. Speech samples were collected from patients in each language, also healthy control speech was included in the dataset. All speakers read a short passage "The North Wind and The Sun" in both languages. Patients in the Hungarian samples were all native speakers, the recordings were done at the Head and Neck surgery department at National Institute of Oncology, and all patients agreed to record their voices for the experiments. A total of 179 recordings were used for the Hungarian patients diagnosed with dysphonia (81 females and 98 males). Alongside these numbers, also 179 healthy control speech samples were included in the Hungarian dataset. The severity level of the dysphonic patients is determined by RBH scale which stands for Roughness, Breathiness and Hoarseness (Schönweiler et al., 2000) with a number between 0 to 3, with 0 specified for the healthy control samples. For Hungarian samples, H is selected as the severity level ranges from 0 to 3, 0 indicates no hoarseness (healthy speaker).

For the Dutch dataset, 30 samples from dysphonic patients have been recorded reading the same "The North Wind and The Sun" passage in Dutch. The recordings were organized at the university hospital of KU Leuven, Belgium. Severities were measured by GRBAS (Grade – overall judgement of hoarseness, Roughness, Breathiness, Asthenia, and Strain), with values ranging from 0 which indicates no hoarseness to 3 means severe dysphonia (Wood et al., 2014). 30 samples from normal-speaking persons were included in the database with the same age distribution as dysphonic samples. Distribution of the severity is shown in Table 1.

Table 1. Severity distribution of samples in datasets

Datasets \ Severity level	0	1	2	3
Hungarian	179	58	63	58
Dutch	30	10	17	3

3.2 Embedding techniques

In this study, two state of the art speaker embedding models have been adopted for the task of binary classification of dysphonic speech and estimating severity level. The X-vector model, which is based on deep neural network architecture, primarily used for speaker verification, has been used for extracting the embedding of the speech samples (Snyder et al., 2018). The architecture is based on the work in (Snyder et al., 2017) with adding data augmentation techniques. The first five hidden layers operate at the frame level using a time-delay neural network. For each time frame ***t*** small temporal context before and after is added in the first three layers. The remaining two layers also work at the segment level but without temporal context, in total frame-level part of the model has a temporal context of 16 frames. The mean and standard deviation of the output of the last layer in the frame level is calculated by the statistic pooling layer. The output of the statistics pooling layer will be used as an input for two hidden layers of size 512 and 300 dimensions. The embedding of 512 dimensions can be extracted for both dysphonic and normal speech from the layer after statistics pooling.

ECAPA-TDNN is based on the X-vector model, extending it mainly in three parts: channel and context dependent statistics pooling, which extends temporal attention to channel attention (Desplanques et al., 2020). This enables the network to focus more on speaker properties and not activate on identical or similar time instances. 1-Dimension Squeeze and Excitation proposed for scaling the frame level features that were limited to 15 frames in the original X-vector to give global properties of the recording. Squeeze and Excitation has been used for computer vision task for modelling global channel interdependencies, also residual block has been used for concatenating 1-Dimentaionl SE to the X-vector model for the sake of keeping the number of parameters relatively near to the original X-vector.

We have downloaded the two pre-trained models using Hugging Face repository[1][2] and SpeechBrain pretrained class. SpeechBrain is an open-source toolkit based on PyTorch. It supports many speech processing tasks such as speech recognition, speaker verification, source separation and many others. The two pretrained models can be accessed and downloaded in the mentioned repository (Ravanelli et al, 2021).

3.3 Classification and regression

Binary classification of dysphonic vs healthy samples has been carried out using SVM (Support Vector Machine) (Cortes et al., 1995) with both linear and rbf kernels. It shows good performance compared to other algorithms such as Decision Tree and K-nearest neighbour with good generalization ability (Dankovičová et al., 2018). Due to the size of datasets and nature of Deep Neural Network (DNN) which requires considerably larger datasets to be able to make a good generalization, using DNN might not be a good for the problem. Classification is done in a cross-lingual approach; first, models were trained on Hungarian samples and tested on Dutch samples, the other direction

[1] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[2] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

was also performed. For both scenarios, hyperparameter tuning was performed for parameters cost and gamma using grid search method with 10-fold cross-validation on the training language. Also, to provide a more accurate assessment of performance, the results of this study were obtained through the implementation of 10-fold cross-validation. This approach allows for a more realistic evaluation of the model's ability to generalize, particularly when the number of training samples is limited.

Estimating the severity level of dysphonic patients has been carried out using Support Vector Regressor (SVR) in both cross-language directions, linear and rbf kernel have been used in the experiments. Also, using the same techniques as in classification phase hyperparameter tuning of both cost and gamma has been performed with 10-fold cross-validation for the training language.

For both classification and severity estimation output features from X-vector and ECAPA were normalized using L2 norm.

3.4 Evaluation metrics

For evaluating binary classification model, we use accuracy, sensitivity, specificity, and F1-score in addition to area under the curve (AUC). Performance of severity estimation of dysphonic patients has been conducted using Root Mean Square Error (RMSE), Pearson correlation and Spearman correlation indicating the linear relationship between actual and predicted severity scores.

4 Results

Binary classification and severity level estimation have been carried out using X-vector and ECAPA-TDNN embeddings extracted as features from speech for both Hungarian and Dutch datasets. The experiments were carried out in cross-lingual nature, where algorithms were trained with speech features from one language and tested using features from the other language.

4.1 Binary classification

Results from binary classification of dysphonic and normal speech are shown in Tables 2 and 3 using embeddings extracted from both models. In the direction where the algorithm trained on Hungarian samples and tested on Dutch utterances. Overall, ECAPA-TDNN model consistently outperformed the X-vector model, achieving an approximately 10% improvement in both accuracy and AUC, particularly when using the rbf kernel. In X-vector scenario linear kernel performs slightly better than rbf kernel in accuracy and AUC. Comparing to the results previously published by (Sztahó et al., 2022); in Hungarian to Dutch direction, it can be concluded that embedding techniques especially ECAPA-DTNN is competitive to the manual feature extraction techniques.

Table 2: Binary classifications results from embedding (Train= Hungarian, Test=Dutch).

	Model	Accuracy %	Specificity %	Sensitivity %	F1-score %	AUC %
X-vector Embedding	SVM-Linear	78.3	74.2	84.0	80.0	79.1
	SVM-rbf	75.0	74.1	75.8	75.4	75.0
ECAPA Embedding	SVM-Linear	83.3	81.2	85.7	83.8	83.4
	SVM-rbf	**85.0**	**92.0**	80.0	**83.6**	**86.0**

In Dutch to Hungarian direction, compared with results from manual feature extraction in binary classification, results reported in the second part of Table 4. It's clear that both embedding methods perform much better than manual feature extraction despite a limited number of speech samples, the best accuracy achieved in manual feature extraction was 66.8% with vowel \E\ included, while our results using embedding methods achieved 86.8% accuracy in case of using rbf kernel which is 20% increase compared to knowledge-based feature engineering.

Table 3: Binary classifications results using embedding (Train= Dutch, Test= Hungarian).

	Model	Accuracy %	Specificity %	Sensitivity %	F1-score %	AUC %
X-vector Embedding	SVM-Linear	75.13	80.0	71.6	72.9	75.8
	SVM-rbf	74.8	79.8	71.2	72.5	75.8
ECAPA Embedding	SVM-Linear	82.6	79.3	86.7	83.5	83.0
	SVM-rbf	**86.8**	**84.3**	**89.7**	**87.3**	**83.0**

Table 4: Binary classifications results from manual feature selection, taken from (Sztahó et al., 2022).

Features		Accuracy %	Specificity %	Sensitivity %	F1-score %	AUC %
Train: Hun	With \E\	**86.2**	**86.7**	**85.7**	**85.7**	**95.5**
Test: Du	Without \E\	81.4	80.0	82.8	81.4	91.0
Train: Du	With \E\	51.6	99.4	2.60	5.00	64.6
Test: Hun	Without \E\	**66.8**	61.9	71.8	68.1	74.3

4.2 Severity estimation

Similarly, predicting the severity level of dysphonic speech has been performed in a multi-lingual fashion. The model was trained using extracted features from Hungarian speech, and tested on Dutch samples, the other way round also performed. The performance of the model is measured using Spearman correlation, Pearson correlation and RMSE. The best results in (Sztahó et al., 2022) by manually extracted features were bolded and underlined in Table 5. The last part of Table 5 refers to the results achieved in Dutch to Hungarian direction using knowledge-based feature extractions, the performance of the model is worse due to the limited number of speech samples in the Dutch database, which made the model unable to make a good generalization.

Table 5: Severity approximation results from manual feature selection, taken from (Sztahó et al., 2022).

	Features	Spearman	Pearson	RMSE
Train = Hu	With \E\	**0.740**	**0.720**	**0.790**
Test = Du	Without \E\	0.660	0.650	0.880
Train = Du	With \E\	0.354	0.382	1.055
Test= Hun	Without \E\	**0.535**	**0.590**	**0.960**

As can be seen from Table 6, our model using embedding features extracted using both speaker verification models outperform hand-crafted feature selection for predicting severity level when we trained the model on Hungarian samples and tested it on Dutch speech, which shows better generalization in all three-evaluation metrics. The best performance was achieved using ECAPA-TDNN model with rbf kernel. Figure 1 shows the scatter plots of original and the estimated H scores for the embedding features, colours indicate the severity scores.

Table 6: Severity approximation results from embedding (Train=Hungarian, Test = Dutch)

	Model	Spearman	Pearson	RMSE
X-vector Embedding	SVR-Linear	0.606	0.638	0.823
	SVR-rbf	0.644	0.677	0.760
ECAPA Embedding	SVR-Linear	0.768	0.770	0.674
	SVR-rbf	**0.769**	**0.771**	**0.650**

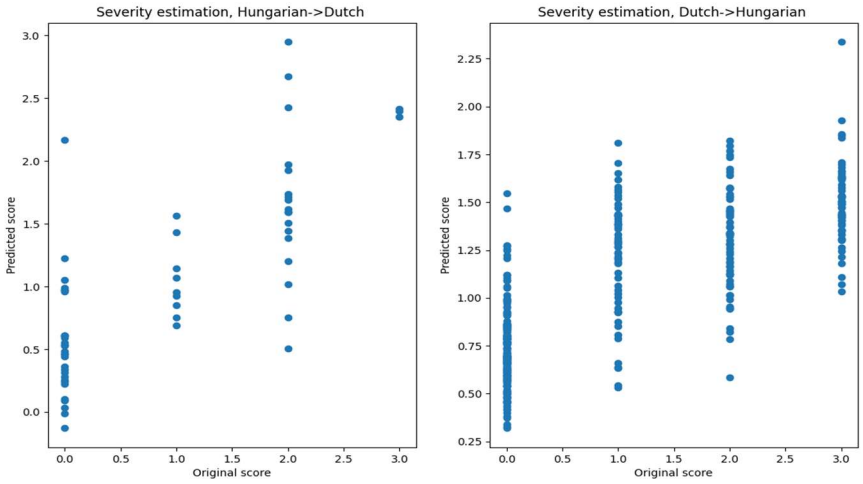

Figure 1: Scatter plot of original and the estimated H scores for both cross-language directions.

Our results from Dutch to Hungarian direction of predicting hoarseness levels is reported in Table 7. Compared to the result from handcrafted features included in Table 5, using features extracted from speaker embedding models performs much better. The capability of ECAPA-TDNN can be seen in Table 7 and scatter plot in Figure 1. Left side of the scatter plot shows severity prediction from Hungarian to Dutch direction. It's clear from the plot the algorithm made better generalizations compared to the right

side of the plot. for estimating the severity level of dysphonic patients in cross-lingual approaches that were not possible using hand-crafted acoustic feature extractions because of the weak performance in predicting severity level.

Table 7: Severity approximation results from embedding (Train=Dutch, Test = Hungarian)

	Model	Spears	Pearson	RMSE
X-vector Embedding	SVR-Linear	0.716	0.662	0.983
	SVR-rbf	0.667	0.599	0.964
ECAPA Embedding	SVR-Linear	0.746	0.730	**0.857**
	SVR-rbf	**0.783**	**0.771**	0.881

5 Discussion

The findings from this experiment indicate that binary classification of normal and dysphonic speech in a cross-lingual manner is possible using deep learning embedding techniques. For the Hungarian to Dutch direction, it can be noted that our results are competitive with manually extracted acoustic features. The best accuracy achieved using manual feature extraction was 86%, in contrast using speaker embedding we achieved 85% accuracy with ECAPA-TDNN.

The biggest difference between our model and model based on knowledge-based features was in Dutch to Hungarian direction. Because of the limited number of samples in Dutch language, training the model with acoustic features and testing on Hungarian samples were not reasonably possible. The best accuracy achieved was 66.8%, while using features extracted from both embedding models, we were able to achieve 75% and 86% accuracy using X-vector and ECAPA-TDNN, respectively.

Embedding techniques adapted from speaker recognition have shown a good result in estimating the severity of dysphonic speech. The two models outperform traditional hand-crafted acoustic features in a cross-language setup (Hungarian to Dutch and Dutch to Hungarian). The severity estimation was performed using Support Vector Regressor (SVR) with both linear and rbf kernels. Nearly 15% difference can be observed in RMSE metrics in Hungarian to Dutch direction (knowledge-based achieved 0.79 while our model achieved 0.65 with ECAPA-TDNN embedding). The highest Pearson correlation achieved was 0.771 in both directions of the model using rbf kernel. Embedding methods achieved 0.796 and 0.783 in Spearman correlation for both Hungarian to Dutch and Dutch to Hungarian direction, respectively. Which is superior compared to manual feature extraction. Due to the size of datasets and nature of Deep Neural Network (DNN) which requires considerably larger dataset to be able to make a good generalization, using DNN might not be a desirable choice for the problem.

The overall observation of this study is that using two state of art speaker embedding extraction methods can be a good replacement for classical hand-crafted acoustic feature extractions for binary classification as well as predicting the severity level of dysphonic speech in a cross-lingual way. The results show that cross-lingual dysphonic speech detection might be possible using deep learning embedding adapted from speaker verification. The models can extract language-independent features from both datasets, and they can be used for the above-mentioned tasks. The main advantage of these embedding models is that they are fast; since they have been trained before, much computation is not required. Another benefit, they don't require hand-crafted feature selection and engineering which is a very problematic and time-consuming task.

Another observation of this study is that ECAPA-TDNN shows better performance compared to the former X-vector in both binary classification and severity level estimation. This is expected because it is an extension of the X-vector, which is the result of the improvements done in former model in three different ways mentioned in section 3.2.

6 Conclusion

In this paper binary classification and severity level estimation of dysphonic speech have been performed in a cross-lingual fashion adapting different speaker verification speaker embedding methods. Two state-of-the-art pre-trained embedding models have been used, which were trained using out-of-domain speech. The embeddings of speech samples in both datasets were extracted using X-vector and ECAPA-TDNN and are used with SVM and SVR for classification and regression. The results proved that these pre-trained deep learning models can be used as a replacement for acoustic feature extraction methods. Models trained on embedding features that are results of pre-trained speaker verification model achieved 85% and 86.6% in binary classification in Hungarian to Dutch, and Dutch to Hungarian, respectively. Compared to manual feature extraction, we can see that our results are competitive in Hungarian to Dutch direction and outperformed in Dutch to Hungarian. This is an indication that we can use these embedding techniques with a small number of samples and still get a good result. Results were superior compared to ones we got from hard-crafted feature engineering. Moreover, the embeddings show a good result in estimating the severity level of dysphonic speech, which leads us to the conclusion that cross-lingual detection is possible with acceptable results.

Acknowledgement

The work was funded by project no. FK128615, which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FK_18 funding scheme. Also, the work was funded by Project no. K128568 with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme.

References

AL-Dhief, F. T., Latiff, N. M. a. A., Malik, N. N. N. A., Sabri, N., Baki, M. M., Albadr, M. A. A., Abbas, A. F., Hussein, Y. M., & Mohammed, M. A. (2020). Voice pathology detection using machine learning technique. 2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT),

Awan, S. N., & Roy, N. (2006). Toward the development of an objective index of dysphonia severity: a four-factor acoustic model. Clinical linguistics & phonetics, 20(1), 35-49.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., & Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. EURASIP Journal on Advances in Signal Processing, 2004(4), 1-22.

Botelho, C., Teixeira, F., Rolland, T., Abad, A., & Trancoso, I. (2020). Pathological speech detection using x-vector embeddings. arXiv preprint arXiv:2003.00864.

Cohen, S. M., Kim, J., Roy, N., Asche, C., & Courey, M. (2012). Prevalence and causes of dysphonia in a large treatment-seeking population. The Laryngoscope, 122(2), 343-348.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Dankovičová, Z., Sovák, D., Drotár, P., & Vokorokos, L. (2018). Machine learning approach to dysphonia detection. Applied Sciences, 8(10), 1927.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Egas-López, J. V., Kiss, G., Sztahó, D., & Gosztolya, G. (2022). Automatic Assessment of the Degree of Clinical Depression from Speech Using X-Vectors. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Grossmann, T., Vaish, A., Franz, J., Schroeder, R., Stoneking, M., & Friederici, A. D. (2013). Emotional Voice Processing: Investigating the Role of Genetic Variation in the Serotonin Transporter across Development. PLoS ONE, 8(7), e68377. https://doi.org/10.1371/journal.pone.0068377

Hadjaidji, E., Korba, M. C. A., & Khelil, K. (2021). Spasmodic Dysphonia Detection Using Machine Learning Classifiers. 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI),

Hegde, S., Shetty, S., Rai, S., & Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. Journal of Voice, 33(6), 947. e911-947. e933.

Lin, H., Karjadi, C., Ang, T. F., Prajakta, J., McManus, C., Alhanai, T. W., Glass, J., & Au, R. (2020). Identification of digital voice biomarkers for cognitive health. Exploration of medicine, 1, 406.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J., Vargas-Bonilla, J., Daqrouq, K., Skodda, S., Rusz, J., & Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. The Journal of the Acoustical Society of America, 139(1), 481-500.

Scheuerer, R., Haderlein, T., Nöth, E., & Bocklet, T. (2021). Applying X-Vectors on Pathological Speech After Larynx Removal. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),

Schönweiler, R., Hess, M., Wübbelt, P., & Ptok, M. (2000). Novel approach to acoustical voice analysis using artificial neural networks. Journal of the Association for Research in Otolaryngology, 1(4), 270-282.

Schwartz, S. R., Cohen, S. M., Dailey, S. H., Rosenfeld, R. M., Deutsch, E. S., Gillespie, M. B., Granieri, E., Hapner, E. R., Kimball, C. E., & Krouse, H. J. (2009). Clinical practice guideline: hoarseness (dysphonia). Otolaryngology–Head and Neck Surgery, 141(1_suppl), 1-31.

Shinohara, S., Omiya, Y., Nakamura, M., Hagiwara, N., Higuchi, M., Mitsuyoshi, S., & Tokuno, S. (2017). Multilingual evaluation of voice disability index using pitch rate. Adv. Sci. Technol. Eng. Syst. J, 2, 765-772.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. Interspeech,

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP),

Stemple, J. C., Roy, N., & Klaben, B. K. (2018). Clinical voice pathology: Theory and management. Plural Publishing.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? Current Opinion in HIV and AIDS, 5(6), 463-466. https://doi.org/10.1097/coh.0b013e32833ed177

Syed, S., Rashid, M., Hussain, S., Imtiaz, A., Abid, H., & Zahid, H. (2021). Inter classifier comparison to detect voice pathologies. Mathematical Biosciences and Engineering, 18(3), 2258-2273.

Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. IEEE circuits and systems magazine, 11(2), 23-61.

Van Houtte, E., Van Lierde, K., & Claeys, S. (2011). Pathophysiology and treatment of muscle tension dysphonia: a review of the current knowledge. Journal of Voice, 25(2), 202-207.

Wood, J. M., Athanasiadis, T., & Allen, J. (2014). Laryngitis. Bmj, 349, g5827. https://doi.org/10.1136/bmj.g5827

Wu, Y., Chen, P., Yao, Y., Ye, X., Xiao, Y., Liao, L., Wu, M., & Chen, J. (2017). Dysphonic voice pattern analysis of patients in Parkinson's disease using minimum interclass probability risk feature selection and bagging ensemble learning methods. Computational and mathematical methods in medicine, 2017.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., ... & Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Sztahó, D., Tulics, M. G., Qi, J., & Vicsi, K. (2022). Cross-lingual detection of dysphonic speech for Dutch and Hungarian datasets. Proceedings Biosignals 2022.