Jurnal Ilmu Sosial dan Pendidikan (JISIP)
Vol. 5 No. 4 November 2021
Terakreditasi Peringkat 5 (No. SK: 85/M/KPT/2020)
e-ISSN : 2656-6753, p-ISSN: 2598-9944
DOI: *10.36312/jisip.v5i4.2279/http://ejournal.mandalanursa.org/index.php/JISIP/index*

Deep Neural Network for Speaker Identification Using Static and Dynamic Prosodic Feature for Spontaneous and Dictated Data

Arifan Rahman
Faculty of Computer Science, Universitas Indonesia

Article Info
Article history:
Diterima: 05 Agustus 2021
Terbit: 02 November 2021

Keywords:
voice
signal processing
deep neural network
prosodic,

Abstract
Kita bisa mengenali seseorang hanya dari suaranya saja. Pada prinsipnya suara memiliki nada (*pitch*) yang berbeda-beda untuk setiap orang. Penelitian ini bertujuan untuk mengukur kinerja Deep Neural Network (DNN) dengan fitur prosodic statis dan dinamis. Prosodik adalah informasi tentang bunyi yang berkaitan dengan nada, intonasi, tekanan, durasi, dan ritme pengucapan seseorang. Data yang digunakan adalah data suara didikte dan spontan yang diambil dari YouTube. Data yang digunakan terdiri dari tiga suara laki-laki dan satu suara perempuan. Data tersegmentasi ke dalam berbagai durasi, 3 detik, 5 detik, dan 10 detik. Setelah data tersegmentasi, fitur prosodik statis akan diekstark dengan 103 dimensi dan fitur prosodik dinamis juga akan diekstrak dengan 13 dimensi. Setiap fitur dan kombinasi fitur akan dilatih dan diuji menggunakan DNN dengan rasio 90:10. Hasilnya menunjukkan bahwa data tersegmentasi 10 detik memiliki akurasi yang lebih tinggi dari yang lain. Akurasi fitur prosodik statis lebih baik daripada fitur prosodik dinamis. Akurasi rata-rata DNN untuk fitur prosodik statis adalah 87,02%. Akurasi rata-rata DNN untuk fitur prosodic dinamis adalah 72,97%. Akurasi rata-rata DNN untuk gabungan fitur prosodik statis dan dinamis adalah 87,72%.

Article Info
Article history:
Accepted: 05 Agustus 2021
Publish : 02 November 2021

Abstract
We can recognize a person by his voice alone. In principle, the sound has a tone (pitch) that is different for each person. This study aims to measure a Deep Neural Network (DNN) performance with static and dynamic prosodic features. Prosodic is information about sound related to tone, intonation, pressure, duration, and rhythm of a person's pronunciation. The data used is dictated and spontaneous voice data that taken from YouTube. The data used consists of three male voices and one female voice. The data is segmented into various duration, 3 seconds, 5 seconds, and 10 seconds. After the data has been segmented, the static prosodic features with 103 dimensions will be extracted and the dynamic prosodic features with 13 dimensions will be extracted too. Each feature and feature combination will be trained and tested using DNN with a ratio of 90:10. The result shows that the 10 seconds segmented data has higher accuracy than the others. Accuracy of static prosodic features is better than dynamic prosodic features. The average accuracy of DNN for static prosodic features is 87.02%. The average accuracy of DNN for dynamic prosodic features is 72.97%. The average accuracy of DNN for combined static and dynamic prosodic features is 87.72%.

Corresponding Author:
Arifan Rahman
Faculty of Computer Science, Universitas Indonesia
Email: arifan.rahman81@ui.ac.id

1. INTRODUCTION

In everyday life, we can recognize other people just by their voices without seeing their faces. Through these sounds, our brains can immediately recognize other people. Each voice has its

characteristics, such as differences between male and female voices, differences of adults and children voices, and other differences. This difference in voice is caused by several factors, such as age, gender, mental and physical conditions [1]. These sound characteristics can be obtained based on the frequency and intensity of the sound source.

In humans, there are two processes of producing sounds which are generation and filtering. Generation is where the sound will be produced in the first place by vibrating the human vocal cords in the larynx and produce periodic sounds. This periodic sound is constant and will be filtered on the vocal track or the articulator, consisting of tongue, teeth, lips, the roof of the mouth, and the others so that the sound becomes an output sound in the form of vowels and consonants.

In principle, the voice consists of several components, namely pitch or tone of voice, formants, and spectrograms, that can be used to identify the characteristics of a person's voice for voice recognition [2]. In this study, researchers will identify human voices using prosodic features and deep neural networks (DNN). Prosodic is information about sound related to tone, intonation, pressure, duration, and rhythm of a person's pronunciation. There are two types of prosodic features, namely static and dynamic.

In this study, we construct a dictated and spontaneous data to measure the performance of each prosodic feature and combination of static and dynamic prosodic feature. The data consisted of four different people, three men, and one woman. Dictated speech is a conversation that is obtained from a recorded human voice that has been prepared in a manuscript. In addition to a more organized speech, environmental conditions can also be arranged so that it is far from noise. Meanwhile, spontaneous speech or spontaneous conversation is a human voice conversation whose narrative is spontaneous without a prepared manuscript. The challenge in using the spontaneous voice dataset is that it finds a lot of interference or noise because the environmental conditions cannot be adjusted. These two features have different levels even though almost have the same representation based on duration, pitch and energy, so that these features are combined in the hidden layer on the Deep Neural Network (DNN). As a comparison, the proposed classification method is compared with Support Vector Machine (SVM) as a baseline. To measure the performance of the model, F1-score are employed.

2. Literature

2.1.Prosodic

Prosodic is information about sound related to tone, intonation, pressure, duration, and rhythm of a person's pronunciation. Prosodic features utilize information about sound frequency related to pitch, intonation, pressure, and energy [3]. The pitch is related to the frequency where a high-frequency sound means it has a high pitch [4]. The pitch is often referred to as the fundamental frequency or F0. Meanwhile, the energy is the level of loudness related to the square of the amplitude [4]. The higher the amplitude value, the higher the loudness value. Figure 1 shows an example of a representation of sound energy based on pitch and amplitude referred on the fundamental value of frequency F0.

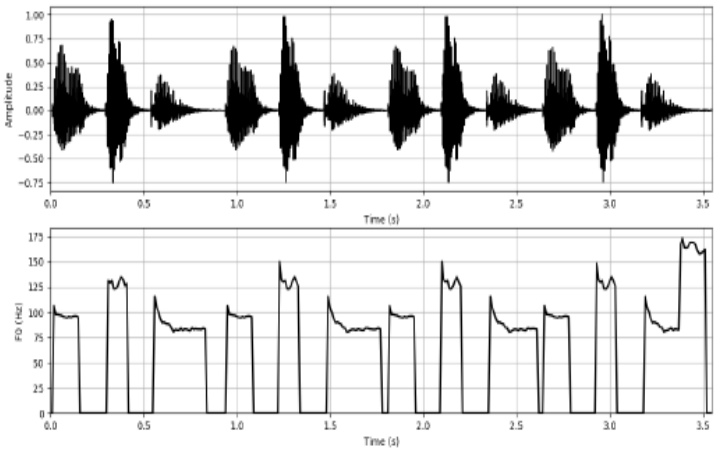

Figure 1. Energy contour and pitch representation on dynamic prosodic

There are two types of prosodic features, namely static and dynamic. Static prosodic features are discrete values obtained from statistical calculations of the basic frequency (F0), energy, or sound duration. The static prosodic feature has 103 features which are divided into three main components as follows.

- Features based on F0.
  - 1-6 F0-contour
  - 7-12 Tilt of a linear estimation of F0 for each voiced segment
  - 13-18 MSE of a linear estimation of F0 for each voiced segment
  - 19-24 F0 on the first voiced segment
  - 25-30 F0 on the last voiced segment
- Features based on energy
  - 31-34 energy-contour for voiced segments
  - 35-38 Tilt of a linear estimation of energy contour for V segments
  - 39-42 MSE of a linear estimation of energy contour for V segment
  - 43-48 energy on the first voiced segment
  - 49-54 energy on the last voiced segment
  - 55-58 energy-contour for unvoiced segments
  - 59-62 Tilt of a linear estimation of energy contour for U segments
  - 63-66 MSE of a linear estimation of energy contour for U segments
  - 67-72 energy on the first unvoiced segment
  - 73-78 energy on the last unvoiced segment Avg., Std., Max., Min., Skewness, Kurtosis
- Features based on duration
  - 79 Voiced rate Number of voiced segments per second
  - 80-85 Duration of Voiced
  - 86-91 Duration of Unvoiced
  - 92-97 Duration of Pauses
  - 98-103 Duration ratios Pause/(Voiced+Unvoiced), Pause/Unvoiced, Unvoiced/(Voiced+Unvoiced), Voiced/(Voiced+Unvoiced), Voiced/Puase, Unvoiced/Pause

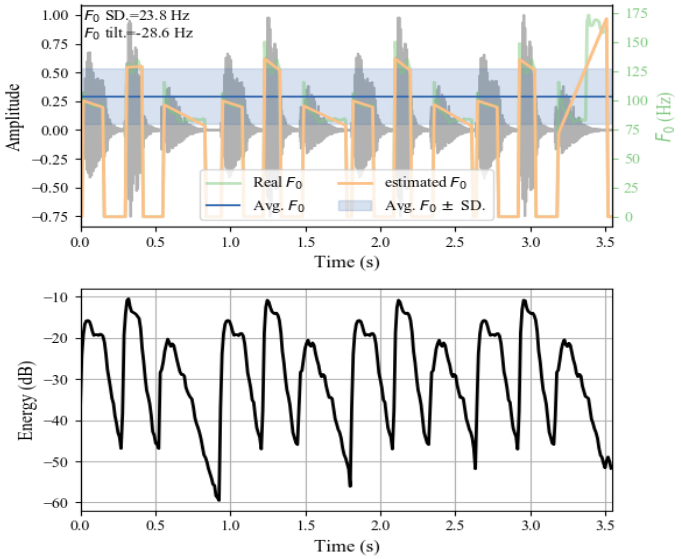

Figure 2. Energy contour and pitch representation on static prosodic

Dynamic prosodic features are continuous values and contours of pitch, energy, or duration, approximated using Legendre polynomials' basic function [5]. The sound unit used in dynamic prosodic features is pseudo-syllables. This truncation of sound to syllable units is based on the valley on the pitch and energy curves. After that, the approximation of the pitch and energy values is carried out with the Legendre Polynomial approximation function.

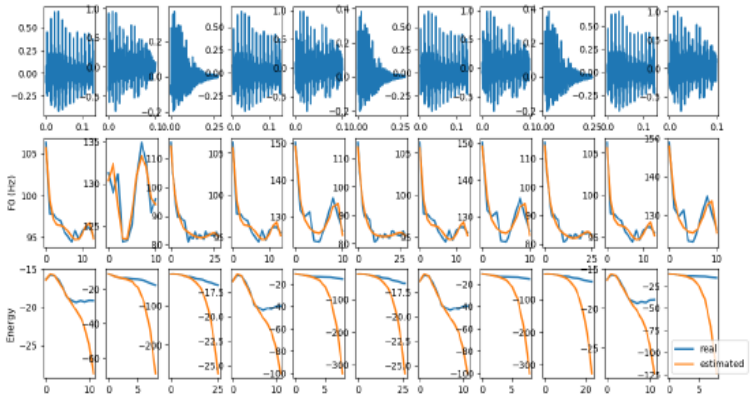

Figure 3. Energy contour and pitch representation on dynamic prosodic

The dynamic prosodic feature has 13 features which are divided into three main components as follows.

- One-dimensional feature duration of sound segmentation. This segmentation is done at the syllable level. The duration value for each syllable will be calculated on a single feature vector.
- Six feature dimensions containing the contour model coefficient of frequency F0. Pitch contour values in each segment are approximated and normalized with Legendre polynomials with a 5-degree coefficient.
  Six feature dimensions containing energy contour coefficients. Energy feature extraction in each segment is approximated and normalized by Legendre polynomials with a 5-degree coefficient as well as the pitch.

2.2.Neural Network and Deep Learning

Neural network is an algorithm that mimics the ability of the human brain to recognize certain patterns. Neural networks process data in a distributed and parallel manner in a unit called an artificial neuron. Neurons are intelligent units that have knowledge through the learning process so that each neuron can make decisions independently and there is a link between one neuron and another. In biological neurons, the input signal is picked up by dendrites which are connected to the axons of other neurons via synapses. The input signal that captured by other neurons is not taken for granted, but there is a weight, then this weight will be added to the weight of the input received. After that, the information is sent to other neurons via the dendrites [6].

Deep Learning is a technique or method in machine learning based on a neural network. The word "deep" in "deep learning" refers to long or deep neural network architecture. The deep learning method can provide the model with the capability to carry out direct learning. Deep learning is possible because it can extract features automatically. Deep learning has several variants or types that are widely used, such as recurrent neural network (RNN), deep neural network (DNN), deep belief network (DBN), and convolutional neural network (CNN).

2.3.Deep Neural Network (DNN)

A deep neural network is a Feedforward Neural Network that can have more than one hidden layer between the input and output layers [7]. The main purpose of a feed-forward network is to approximate the function. DNN consists of three main elements: the input layer, hidden layer, and output layer. A deep neural network is a Feed forward Neural Network that can have more than one hidden layer between the input and output layers [7].

Input layer serves to receive input data. The number of neurons in the input layer is generally the same as the number of features to the network [8]. Hidden layer receives data from the input layer and serves non-linear model functions. The modelling is done by using the activation function in the hidden layer. Several activation functions used in the hidden layer are sigmoid, tanh, hard tanh, and rectified linear (ReLU).

Output layer serves to provide predictive results from the model that has been made. The output can be real-valued for the regression case or a set of probabilities for the classification case. There are two types of activation functions used for the classification process in mapping the output, namely sigmoid and softmax. Sigmoid is used for binary classification, while softmax is used for multiclass classification. Softmax will produce a probability distribution between classes.

In addition to layers, some important terminology in neural network architecture includes data input and targets, loss functions and optimizers [9]. The loss function defines the feedback signal used for the learning process. Meanwhile, the optimizer functions to determine the learning progress.

The network comprises interconnected layers and maps the input data to a prediction class; the loss function compares the model's predictions with the target class and produces a loss score. The loss score serves to measure the predictions made by the model for the target class and uses the optimizer to update the network weight to get the best performance in the training process.

Another essential concept is a generalization. Generalization is the ability of a model to predict data that has never existed before. At the beginning of the training, optimization and generalization are interrelated where the loss value in training data and testing data is equally low. This condition is called underfitting, where the model cannot map the pattern to the training data. But after a few iterations, generalization stops improvising and performance drops. This condition is called overfitting.

To overcome overfitting, one solution that can be done is to add training data. However, if this is not possible, the way to do this is regularization. Some regularization techniques that can be done are as follows.

- Reducing network's size by reducing the layer's size and neurons per layer's size.
- Added weight regularization, where this method works by adding a limit to the complexity of the network by forcing the weights to take only the small value so that the weight distribution becomes more regular. There are two types of regularization, namely L1 regularization and L2 regularization.
- Added dropouts. Dropout works by removing neurons on the layer. The dropout value commonly used is between 0.2 and 0.5 [9].

Implement early stopping. Early stopping works by stopping the training process when performance stops making improvements to the validation set.

Table 1. Type of Activation Function and Loss Function

Classification	Activation Function	Loss Function
biner	Sigmoid	binary crossentropy
multiclass	Softmax	categorical crossentropy
multilable	Sigmoid	binary crossentropy

3. HASIL PENELITIAN DAN PEMBAHASAN

3.1.System Development Environment

The system development environment used in this study is summarized in Table 2.

Table 2. System Development Environment

Type	Specification	About
Operation System	macOS Catalina 10.15.7	
RAM	16GB	
CPU	2,3 GHz Quad-Core Intel Core i7	
GPU	NVIDIA GeForce GT 650M 512 MB Intel HD Graphics 4000 1536 MB	
Softwares	Python 3.7.3	Primary programming language
	Tensorflow 2.5.0	Python library for deep-learning (back-end)
	Keras 2.4.3	Python library for deep-learning
	FFMpeg 4.4	Python library for segmentation
	Praat 6.1.49	Software for prosodic extraction
	Audacity 2.3.3	Software for prep-processing data
	Unsilence 1.0.8	Python library for remove silcnce in audio file

3.2.Data Collection

The data used is dictated and spontaneous voice data that taken from YouTube. The data are the dictated voice data from speech and from interviews of the president, prime minister, vice president and the former president of the Republic of Indonesia. The data used consists of three male voices and one female voice. Data is saved in *.wav* format.

3.3.Pre-Processing Data

Voice data that has been collected will be pre-processed first. Pre-processing of data is carried out by eliminating unimportant segments and homogenizing the duration. Because this study focuses on speaker identification, non-essential segments such as unclear voice noise, unclear word pronunciation, and possible external noise will be eliminated first. The process of removing the noise is done manually using *Audacity* software. It also removes the part that has no sound with the Python library called *Unsilence*. The recorded sound that has been produced from the process of eliminating unimportant segments will be segmented into three different duration variations, namely 3 seconds, 5 seconds, and 10 seconds. Segmentation is done using a Python program with the *FFMpeg* library. The segmented audio file is saved again in *.wav* format. Table 3 shows the data after pre-processing.

Table 3. System Development Environment

	Type	**Duration**	**Total File**		
			3s	**5s**	**10s**
JW	Dictated (D)	11:29	280	168	84
	Spontaneous (S)	13:58	230	138	69
JK	Dictated (D)	13:32	436	262	131
	Spontaneous (S)	21:46	271	163	81
HB	Dictated (D)	10:32	537	322	161
	Spontaneous (S)	26:49	211	127	64
SM	Dictated (D)	20:45	624	375	188
	Spontaneous (S)	31:11	415	249	125
Total			**3004**	**1804**	**903**

3.4.Feature Extraction

Prosodic features were extracted with a python library called *NeuroSpeech* [10]. *NeuroSpeech* uses a Python library called *Praat* to extract prosodic features. There are three main features extracted, namely duration, fundamental frequency (F0) or pitch contour model and energy contour model. For static prosodic, one *.wav* file will produce 1 vector with a length of 103 while for dynamic prosodic it will produce 1 vector with a length of 13 dimensional features.

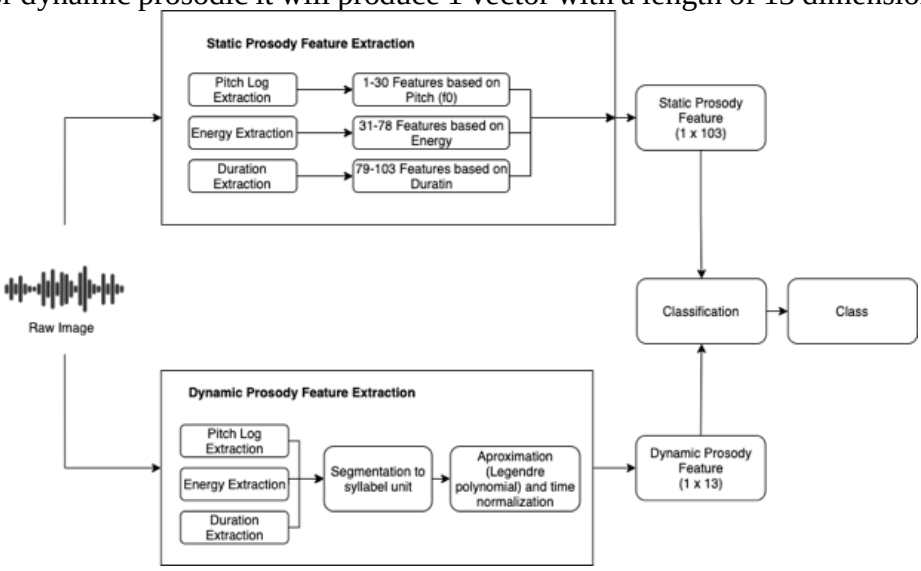

Figure 4. Energy contour and pitch representation on static prosodic

3.5.DNN Model Design

The classification process is implemented using DNN and as a comparison, SVM is also implemented. The formation of the DNN model is carried out with a Python library called *Keras* which uses a *tensorflow back-end*. Meanwhile, SVM is implemented with a linear kernel with Scikit-learn parameters. In the feature combination scenario, individual features are combined directly on the input features before being modeled with DNN.

Static prosodic features have dimensions of 103 features. Meanwhile, dynamic prosodic features have dimensions of 13. Each of these features will be tested using DNN first. The number of hidden layers used is 1. Meanwhile, the number of neurons in the input layer will be used the same as the number of features, namely 103 for static prosodic features and 13 for dynamic prosodic features. Dropout regulariation is not applied to these two models.

For the whole model, the hidden layer uses the Rectified Linear Unit (ReLU) activation function. Meanwhile, because this is a multiclass classification, the output layer uses the softmax activation function and the loss function used is categorical crossentropy. Furthermore, in the DNN model an optimizer function is also applied. In this study, all DNN models were created using the RMSProp optimizer function. The selection of RMSProp was based on recommendations from Goodfellow, 2016 which recommended an algorithm with an adaptive learning rate [11].

For the feature merging scenario, it is done by combining the features in the input layer using the concatenate function. The feature combination consists of combining static prosodic features with dynamic prosodic features. This combination of features has dimensions of 116 features that will become neurons in the input layer. These features also use the Rectified Linear Unit function on the hidden layer and also use the RMSProp optimizer function. For the selection of training and testing data, k-fold cross validation is carried out with k=10 and the training and testing data comparison is 90:10.

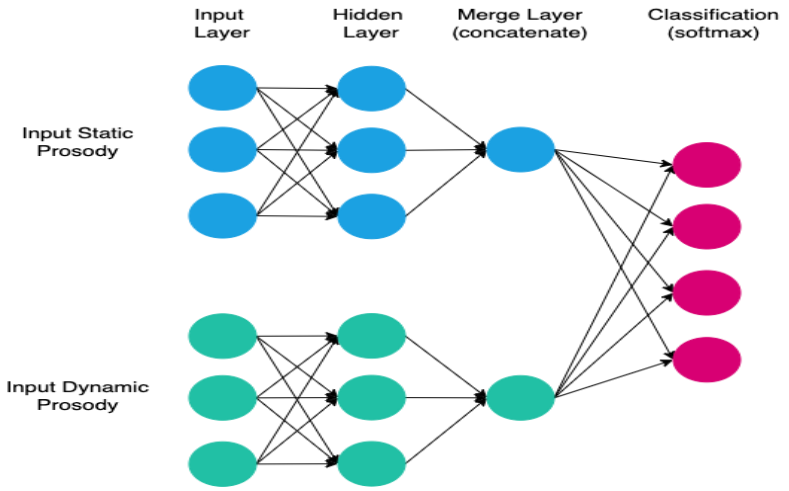

Figure 5. DNN architecture for combination of static and dynamic prosodic feature

3.6.Evaluation

To measure the performance of prosodic feature in speaker identification, an evaluation metric namely F1-score is used. F1-score is a harmonic calculation of precision and recall. In addition, a confusion matrix is employed to understand the level of accuracy and error in each language class. The formulation of the F1-score is defined as follows.

$$\text{F1} - \text{score} = 2.\frac{\text{precission} * \text{recall}}{\text{precission}+\text{recall}} \qquad (1)$$

This study aims to investigate the performance of our proposed features in identifying speaker on spontaneous dan dicated data. Deep Neural Network (DNN) is chosen as a classifier for both individual features and combine features. Support Vector Machine (SVM) is used as a baseline.

4. RESULT

4.1.Result of Static Prosodic

Static prosodic features are discrete values obtained from statistical calculations of the basic frequency (F0), energy, or sound duration. The static prosodic feature has 103 dimension features. The result of experiment on static prosodic feature with 103 dimension features is summarized in Table 4.

Table 4. Experiment result of Static Prosodic Feature

		SVM (%)	DNN (%)
Spontaneous (S)	3s	81.48	85.60
	5s	84.87	84.13
	10s	90.47	90.42
Dictated (D)	3s	84.74	85.80
	5s	87.50	87.09
	10s	92.50	**94.67**
S + D	3s	79.15	82.71
	5s	85.34	84.88
	10s	87.37	87.89
Average		85.94	**87.02**

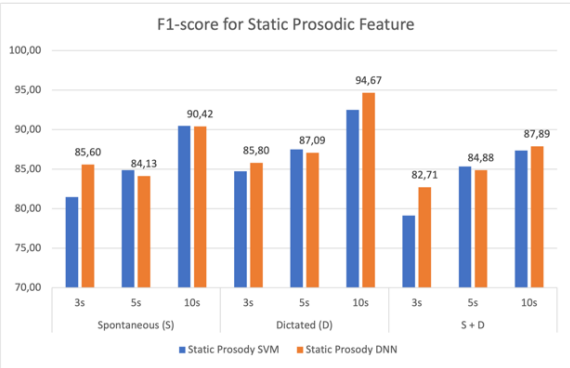

Figure 6. Graphic experiment result of Static Prosodic Feature

Based on Table 4, it can be seen that DNN implementation on the static prosodic feature resulted in the highest f1-score of 94.67% for 10 seconds of dictated data. In general, DNN has a higher average of f1-score value compared to SVM classification. The data also shows that the accuracy of the 10s data is always higher than the 3s data and 5s data.

4.2.Result of Dynamic Prosodic

The result of experiment on dynamic prosodic feature with 13 dimension features is summarized in Table 5.

Table 5. Experiment result of Dynamic Prosodic Feature

		SVM (%)	DNN (%)
Spontaneous (S)	3s	64.02	64.73
	5s	74.78	71.73

	10s	**87.30**	78.44
Dictated (D)	3s	72.88	75.03
	5s	80.55	74.01
	10s	80.00	80.00
S + D	3s	61.23	61.92
	5s	73.82	74.69
	10s	86.40	76.15
Average		**85.94**	**75.66**

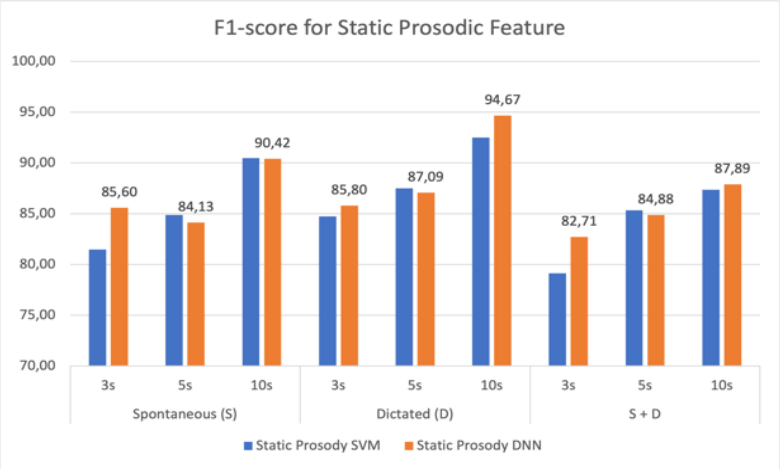

Figure 7. Graphic experiment result of Dynamic Prosodic Feature

Based on Table 5, it can be seen that SVM implementation on the static prosodic feature resulted in the highest f1-score of 87.30% for 10 seconds of spontaneous data. In general, SVM has a higher average of f1-score value compared to DNN classification. The data also shows that the accuracy of the 10s data is always higher than the 3s data and 5s data.

4.3.Result of Merging Static and Dynamic Prosodic

The result of experiment on merging the static and dynamic prosodic feature with 116 dimension features is summarized in Table 6.

Table 6. Experiment result on Combination of Static and Dynamic Prosodic Feature

		SVM (%)	DNN (%)
Spontaneous (S)	3s	78.30	87,46
	5s	89.07	89.38
	10s	93.65	91.02
Dictated (D)	3s	84.74	87.01
	5s	87.50	87.91
	10s	92.50	**93.67**
S + D	3s	77.85	80.26
	5s	84.29	85.73
	10s	88.34	87.02
Average		**85.94**	**86.25**

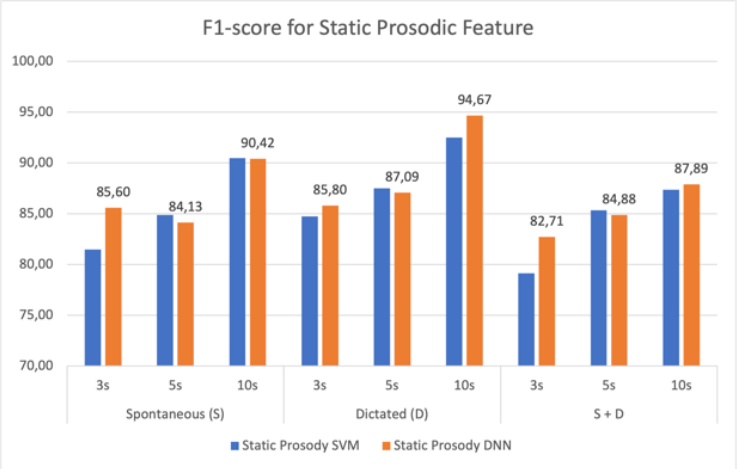

Figure 8. Graphic experiment result on Combination of Static and Dynamic Prosodic Feature

Based on Table 6, it can be seen that DNN implementation on the static prosodic feature resulted in the highest f1-score of 93.67% for 10 seconds of dictated data. In general, DNN has a higher average of f1-score value compared to SVM classification. The data also shows that the accuracy of the 10s data is always higher than the 3s data and 5s data.

5. CONCLUSION

This study aims to measure the performance of the Deep Neural Network to identify the human voice using static and dynamic prosodic features. Static feature for 103 dimensions and dynamic feature for 13 dimensions. Prosodic is information about sound related to tone, intonation, stress, duration, and rhythm of a person's pronunciation. The prosodic feature utilizes information about the frequency of the sound in terms of pitch, intonation, stress, and energy.

The classification process used DNN with 13 neurons in the input layer for dynamic prosodic feature, 103 neurons for static prosodic feature and 116 neurons for combination of these features, one hidden layer. The hidden layer used is the Rectified Linear Unit (ReLU) activation function. Meanwhile, the output layer uses an activation function and a loss function suitable for multiclass classification. In this case, the activation function used in the output layer is softmax, and the loss function is categorical cross-entropy.

Furthermore, in the DNN model, the optimizer function is also used. In this study, the DNN model was built using RMSProp as the optimizer function. The selection of the RMSProp was based on the recommendations of Goodfellow et al. (2016), who recommend algorithms with adaptive learning rates (in this case, RMSProp and AdaDelta) because they have robust performance [11].

The result shows that the 10 seconds segmented data has higher accuracy than the others. This is because the smaller the segmentation of the data, then there will be parts of the word that are segmented incompletely so that the feature extraction data of a smaller duration is less described.

Accuracy of static prosodic features is better than dynamic prosodic features. The accuracy of static prosodic features is better than dynamic prosodic features. This is because the features presented in the static prosodic features are more and more complete. The average accuracy of DNN for static prosodic features is 87.02%. The average accuracy of DNN for dynamic prosodic features is 72.97%. The average accuracy of DNN for combined static and dynamic prosodic features is 87.72%.

Based on Figures 6, 7 and 8, the performance results from DNN do not always outperform the performance results from SVM, this is because there is not too much data to process DNN, because DNN will achieve good performance if it has a lot of training data. In addition, combining features will give better results if only using each feature

In this study, the authors realize that there are still many shortcomings in the study. Future research can add more types of voice variations and implement segmentation techniques automatically, for example, by implementing Voice Activity Detection (VAD). In this research, voice segmentation is still done manually. For further research, it is recommended that voice-by-speech segmentation be segmented automatically, for example, with the Voice Activity Detection technique. And then add more data to identify human voice for DNN because the more data, the better the performance of DNN.

6. REFERENCES

[1] Kirkov, B., & Zielinski, T. P. (2019). Formant Analysis of Traditional Bulgarian Singing from Rhodope Region. *Signal Processing - Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, SPA*, *2019-Septe*, 148–152. https://doi.org/10.23919/SPA.2019.8936714

[2] Wicaksono, G., & Prayudi, Y. (2013). Teknik Forensika Audio Untuk Analisa Suara Pada Barang Bukti Digital. *Semnas Unjani*, *December 2013*.

[3] Li, H., Ma, B., & Lee, K. A. (2013). Spoken language recognition: From fundamentals to practice. *Proceedings of the IEEE*, *101*(5), 1136–1159. https://doi.org/10.1109/JPROC.2012.2237151

[4] Jurafsky, D., & Martin, J. H. (2009). Speech and language processing. In *Day-to-Day Dyslexia in the Classroom* (2nd ed.). Prentice-Hall, Inc. https://doi.org/10.4324/9780203461891-3

[5] Dehak, N., Dumouchel, P., & Kenny, P. (2007). Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(7), 2095–2103. https://doi.org/10.1109/TASL.2007.902758

[6] Zhang, V. J., Shao, L., & Zhang, L. (n.d.). *Igor Aizenberg Complex-Valued Neural Networks with Multi-Valued Neurons Studies in Computational Intelligence*. *353*.

[7] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, *29*(6), 16–17. https://doi.org/10.1109/MSP.2012.2209906

[8] Patterson, A. J., & Gibson, A. (2019). Deep Learning: A Practitioner's Approach. In *O'Reilly Media, Inc*.

[9] Chollet, F. (2018). Deep Learning with Python. In *2018 21st International Conference on Information Fusion, FUSION 2018*. Manning Publications Co. https://doi.org/10.23919/ICIF.2018.8455530

[10] Orozco-Arroyave, J. R., Vásquez-Correa, J. C., Vargas-Bonilla, J. F., Arora, R., Dehak, N., Nidadavolu, P. S., Christensen, H., Rudzicz, F., Yancheva, M., Chinaei, H., Vann, A., Vogler, N., Bocklet, T., Cernak, M., Hannink, J., & Nöth, E. (2018). NeuroSpeech: An open-source software for Parkinson's speech analysis. *Digital Signal Processing: A Review Journal*, *77*, 207–221. https://doi.org/10.1016/j.dsp.2017.07.004

[11] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. 1–3. http://www.deeplearningbook.org/front_matter.pdf