INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Vocalization Recognition of People with Profound Intellectual and Multiple Disabilities (PIMD) Using Machine Learning Algorithms

Waldemar Jęśko

Poznan Supercomputing and Networking Center, Poznan, Poland
waldemarj@man.poznan.pl

Abstract

We investigate vocalization recognition for people with Profound Intellectual and Multiple Disabilities using various machine learning algorithms. The amount of training data available for people with PIMD is typically significantly limited. Due to this fact, data augmentation process was used. Various types of Machine Learning algorithms were tested: k-NN, NB, DT, RDF, MLP and LSTM. During research we also tested various regularization techniques to improve recognition performance. The best results were obtained in case of MLP network with dropout and batch normalization: 90%.

Index Terms: speech recognition, vocalization recognition, PIMD, Profound Intellectual and Multiple Disabilities, AI

1. Introduction

Machine learning (ML) is widely used in many fields including medicine, autonomous vehicles, industry, logistics and speech recognition. Over the years, many methods have been developed to increase the accuracy of speech recognition. The method that achieves the best results leverages artificial neural networks (ANN). In this paper, we investigate applicability of neural networks in recognition of vocalization of people with Profound Intellectual and Multiple Disabilities (PIMD).

The purpose of our work on vocalization recognition is to support an ICT platform implemented for INSENSION [1] project. The platform's objective is to improve quality of life of people with PIMD [1]. Recognizing vocalizations of a PIMD person, can help to determine the emotional state of such a person. Emotions expressed by vocalizations of PIMD people are easily recognizable by their caregivers but not by other people who are not familiar with their behavior. The mechanisms implemented in the INSENSION platform leverage several technologies such as image processing for facial expression and gesture recognition, and vocalization recognition. This approach enables better interpretation of the current state and needs communicated by a PIMD person and consequently also enables the system to automatically control parameters of the surrounding environment, e.g. adjust room temperature.

Vocalizations are non-linguistic sound events that are not related to any form of speech. They include sounds such as laughing, crying, breathing, etc. In the case of people with PIMD, vocalizations are the only form of audio communication. In addition to the basic (i.e. typical) types of vocalizations, PIMD people additionally vocalize various (generally undefined) sounds trying to convey their current emotions. It is worth noting that, unlike typical vocalizations and speech, the vocalizations of people with PIMD are quite individual and unique in terms of phonetics. Therefore, vocalization recognition model must be trained for each person separately. Due to the very individual character of vocalizations it is impossible to use advanced automatic speech recognition systems.

In this paper we present work on recognition of vocalizations of PIMD people. We investigate the use of various ML algorithms, such as k-nearest neighbours (k-NN), naive-bayes (NB), decision trees (DT) or neural networks (NN) and identify problems that affect the results. The recognition results quality is assessed in the context of the INSENSION platform and its objective to support PIMD people. The paper is organized as follows. In section 2 we present state-of-the-art in the field of vocalization recognition. Section 3 describes experiments with machine learning based recognizer, experimental results and their analysis. Conclusions and future work directions are presented in Sections 4 and 5.

2. Related Work

Recognition of typical non-linguistic vocalizations has been carried out using many different machine learning techniques. They include basic methods such as k-NN, decision trees, support vector machine (SVM)[2, 3] and more advanced solutions based on various types of artificial neural networks. Recognition of single typical vocalizations such as laugh [4], cry [5] or cough [6], and of multiple vocalizations [3, 7, 8], have been studied. For the most part, neural network architecture used for recognition was a recurrent neural network, or more precisely LSTM network (Long Short-Term Memory) [6, 7, 8, 9, 10]. These networks are very good at modeling temporal aspect of data. LSTM cells model long-term time relationships between data, making them suitable for speech recognition tasks.

In some cases information used in vocalization recognition came from two independent sources: audio and video. According to the results presented in [7, 8], this approach increases the accuracy of typical vocalizations recognition. The disadvantage of this approach is the necessity to obtain an additional data stream, process data and extract relevant information. Predictions generated by a vocalization recognition described in this paper will be combined with the results of another INSENSION platform module to obtain a more reliable assessment of a person with PIMD emotional state. However, in this paper we focus on the vocalization recognition results alone.

It is also worth noting that the vast majority of the studies related to the recognition of vocalization concerned animal

[1] https://www.insension.eu

http://dx.doi.org/10.21437/Interspeech.2021-1239

vocalization [11, 12, 13, 14, 15, 16]. In this case there are no disturbances in the form of normal speech whose presence in the recordings significantly affects the quality of vocalization recognition for people. Despite the fact that these are different vocalizations, more recent studies in the field of animal vocalization recognition [14, 16] show that LSTM-based models achieve one of the better recognition results reported (accuracy above 90%). In the above-mentioned studies high recognition accuracy results can be attributed mainly to a large amount of training data used (several hundred or several thousand representations).

In contrast to the previous publications, our research is related to the vocalizations of people with PIMD that are significantly individual and specific. People with PIMD are at an early stage of speech development. Their vocalizations, which are a form of communication, resemble sounds generated by healthy children at an early stage of their development, i.e. the babbling stage - heard sounds resembling ta, ma, ba, la syllables, where articulation is controlled (therefore they can be recognized). Apart from the additional general sounds (crying, laughter, coughing, "breathing" sounds), people with PIMD also very frequent vocalize sounds resembling the vowels aaa, aeaeae, eee and the syllables ge, ga, echo, ah, etc. These vocalizations correspond much more to the natural language and are more individual than other vocalizations - animal vocalizations or the sounds of crying, laughing or coughing. Additionally, compared to typical animal or human vocalizations, obtaining the sufficient amount of training data to recognize the vocalization of people with PIMD is much more difficult.

In order to implement the vocalization recognition module for people with PIMD, we have decided to investigate the possibility of using various ML algorithms and to apply the most efficient of them to the INSENSION system.

3. Experiments and results

In order to carry out the research, a vocalizations database was prepared using recordings acquired in the INSENSION project for two children with PIMD. The obtained data made it possible to conduct independent studies for two people with PIMD and to create models of vocalization recognition for the target environment in which the INSENSION system will be implemented.

3.1. Dataset Preparation

Annotation was carried out with the ELAN: Linguistic Annotator program [17]. During annotation of the recordings, not only vocalizations but also other states were marked, e.g. communication attempt, inner state, etc. (in the presented research, the focus was only on recognizing vocalization). Each recording was annotated by one of the caregivers of the person with PIMD, who knew the different mental states of the person under his or her care and was able to distinguish vocalizations and the corresponding states. Only these people were able to mark the appropriate states and vocalizations. In the next step an additional annotation verification process (including accuracy verification) was carried out by another person who also looked after and knew the person with PIMD to whom the given recording was related.

At this stage of the project, it was possible to obtain recordings only for 2 people with PIMD, who we refer to as Person A and Person B. It should be emphasized that recording sessions of people with PIMD are very time-consuming due to the specific environmental conditions that must be met (the possibility of recording only in special, adapted rooms of care facilities where people with PIMD spend only a limited time) and the need for caregivers' participation in the process of recording. An additional difficulty during the recording process is the fact that people with PIMD (due to their limitations) do not cooperate especially in this recording task. It also influences the amount and type of vocalized sounds that cannot be forced or provoked (especially vocalizations related to negative events / feelings) by the caregivers in order to obtain specific, missing vocalization representations. The above-mentioned difficulties make acquiring recordings of people with PIMD a challenging task.

Due to the diversity and specificity of PIMD vocalization, recognition models must be created and trained for each person separately. Based on the data extracted from the recordings a list of classes for each person that will be recognized by the target ML model was created. The number of instances for each identified vocalizations (classes) in the obtained recordings is presented in Table 1.

Table 1: *Number of annotated vocalization instances.*

Person	**Vocalization examples**					
Person A	eee	grunt	aaa	moan	eeh	laugh
	83	24	23	20	10	6
Person B	aaa	crying	aeaeae	cough	eee	nge
	133	51	22	11	6	4

The first step in preparing the learning data set was the parameterization of the speech signal. In this process, a raw speech signal is converted into vectors of numerical values representing only relevant information. A standard speech signal parameterization method was used: MFCC [18] (25 ms frame, 10 ms step, 40 features). In the target system, vocalizations will be recognized in an uninterrupted audio stream (continuous recordings). Therefore, in the next pre-processing step, data extraction through an intermediate buffer of a specific length has been implemented to enable sequential buffering of the data, which is extracted from the audio stream. In the second step the extracted features were divided into intermediate windows using an additional buffer. During processing this buffer is used to extract data from the recording for a given time period and to pass it to the recognition model in order to detect events. Additional processing is necessary to recognize vocalizations in continuous recordings where there is no information about the beginning and the end of a given audio event. For this reason, the extracted data had to be processed in the next step of division into examples with a defined buffer length. The length of this buffer was one of the parameters investigated during the model tests. The tested length values were selected experimentally based on the length of the vocalizations present in the recordings. The smallest buffer length (50 ms) was selected as 25% of the shortest vocalization duration. The largest value (1000 ms), is the most common duration of vocalizations in the set. The number of vocalization instances in relation to the buffer size for Person A is presented in Table 2 (the number of representations for Person B was similar). The number of examples for a particular vocalization (Table 2) depends strictly on the duration of the vocalization for which these instances were created. The longer the vocalization, the more time frames can be extracted.

Table 2: *Number of samples for tested buffer sizes – Person A.*

Buffer length [ms]	Vocalization examples					
	eee	grunt	aaa	moan	eeh	laugh
50	2198	749	832	901	188	136
100	1080	368	410	445	92	67
200	518	178	198	218	43	33
300	335	114	129	142	28	21
500	182	65	73	80	15	12
1000	65	25	31	34	3	4

After the initial data processing, preliminary research closely related to data preparation was carried out. These studies were performed in order to select the most optimal form of pre-processed data. Various configurations of the following parameters were tested: buffer length, buffer overlapping, data augmentation. In the initial research several machine learning algorithms were used. In the future studies they will be tested in more details using the selected, most optimal form of data. Finally, data processed using the following settings were selected for further experiments:

- buffer length: 500 ms,
- buffer overlapping: No,
- data augmentation: Yes (4-fold increase).

In the augmentation process, various noises were used, which were imposed on the original recordings. In addition, the data set was increased by changing various recording parameters (e.g. bandpass, volume). Augmentation was carried out at the level of input audio files, the number of which was increased 4 times. The data set created with the selected data processing configuration for learning ML algorithms is presented in Table 3.

Table 3: *Number of examples in the final dataset – 500 ms buffer length, no overlapping, 4-fold data augm.*

Person	Vocalization examples					
Person A	eee	grunt	aaa	moan	eeh	laugh
	728	260	292	320	60	48
Person B	aaa	crying	aeaeae	cough	eee	nge
	1928	2884	248	56	60	20

Furthermore, an additional class was included in the final dataset: "bck" (background). Background class is related to all other audio signals that are present in recordings: noises, silence, caregiver speech etc. During the training process, all parts of the recordings that were not marked as vocalization were used to train the model to recognize and distinguish the additional background class. After the augmentation process, the background class was the most represented class (about 60K examples). Therefore, the vocalization objects were additionally duplicated to equalize the number of objects to background objects number in the training set. The equalization of the number of class objects in the training set is important for the model to learn each class equally well in each training epoch and not to adapt only to specific, more representative classes. Thus, the data duplication process which was used matched the number of each class objects in the training to the number of objects in the most represented class. Finally, the tested ML models recognizes the following classes:

- Person A (7): eee, grunt, aaa, moan, eeh, laugh, bck.
- Person B (7): aaa, crying, aeaeae, cough, eee, nge, bck.

3.2. Experimental Results

The research was divided into 3 stages. In the first step, a dummy classifier was used to generate a baselines result. In the next stage, basic machine learning algorithms (e.g. k-NN, DT etc.) were tested. In the last phase, the efficiency of feedforward and recurrent neural networks applied in the analyzed research problem was examined.

Due to the amount of data which was relatively small for a machine learning task, the tests were carried out using 5-fold cross-validation. In order to maintain the same class distribution in each subset, a special type of validation was used: stratified k-fold cross validation with additional random division of the data set. The data was divided into folds at the level of input audio files, so that indirect instances of a given vocalization did not occur both in the training set and in the test set (data leakage between sets). In all the conducted studies, the following metrics were used to assess the quality of the models: accuracy (acc), precision (prec), recall (rec) and f1-score (f1).

3.2.1. Baseline

In order to generate baseline results, a dummy classifier has been implemented, which was based on three standard statistical / random methods:

- "uniform": generates predictions uniformly at random.
- "stratified": generates predictions by respecting the class distribution of the training set.
- "most frequent": always predicts the most frequent class from the training set.

The dummy classifier results are summarized in Table 4.

Table 4: *Baseline results.*

Person	Dummy method	Metrics [%]			
		acc	prec	rec	f1
Person A	uniform	**18.1**	**30.2**	**18.1**	**21.4**
	stratified	11.2	17.4	11.2	12.8
	most frequent	10.3	1.10	10.3	1.90
Person B	uniform	15.0	32.1	15.0	**20.1**
	stratified	14.9	**33.0**	14.9	19.9
	most frequent	**33.9**	11.5	**33.9**	17.2

3.2.2. Basic machine learning methods

In the next stage of the research, the selected methods of ML were tested in order to verify whether they would be sufficient for the research problem under consideration. The following basic ML algorithms were tested: k-NN, NB, DT and random decision forest (RDF).

The training data in the prepared database is referenced in time (the temporal aspect of the data). On the other hand, the simple machine learning algorithms are adapted to one-dimensional data. Hence, in the pre-processing stage, the data had to be additionally flattened (transformed into a one-dimensional array). For this purpose, two methods were used:

1. data flattening: 2d data to 1d conversion (simple data flattening).

2. data reshape: data transformation by computing the mean value of each MFCC parameter over time.

Investigating many various machine learning methods for which many various parameter values must be tested is quite time-consuming. Therefore, a random search technique was used to test multiple ML method configurations within an

acceptable time period. Table 5 shows the results of the basic machine learning methods generated using data reshape transformation method, for which better results were obtained.

Table 5: *Basic ML methods results.*

Person	Metrics	Basic ML algorithms			
		k-NN	NB	DT	RDF
Person A	acc	**50.9**	49.1	42.2	31.0
	prec	**50.9**	47.7	40.6	25.3
	rec	**50.9**	49.1	42.2	31.0
	f1	46.4	**47.4**	34.5	18.7
Person B	acc	**74.5**	66.9	58.1	58.3
	prec	76.5	**79.8**	61.0	69.4
	rec	**74.5**	66.9	58.1	58.3
	f1	**73.0**	70.9	54.2	53.9

3.2.3. Neural networks

In the final phase of the research, the following neural network architectures were tested: Multi-layer Perceptron (MLP, feedforward NN), Long-Short Term Memory (LSTM, recurrent NN) and hybrid architecture: LSTM with additional dense layers (regular deeply connected NN layers). For each of these networks, a number of tests were carried out in relation to their main parameters: various number of layers, number of neurons in layers, optimizer methods and activation functions.

Additionally, neural networks investigation was extended to include the verification of the influence of various regularization techniques on the final results. The following regularization methods were verified: dropout, weight regularization (weight reg.), batch normalization (batch norm.) and combination of many regularization methods (many methods). Early stopping of the learning process was also used in each test to further reduce the model overfitting. The final results of vocalization recognition for Person A and Person B are presented in Table 6 (average for all 4 metrics).

Table 6: *Neural networks results.*

Person	Reg. method	NN architectures		
		MLP	LSTM	LSTM +dense
Person A	-	**70.5**	69.1	67.4
	dropout	**76.3**	71.7	66.7
	weight reg.	70.9	**72.5**	68.2
	batch norm.	**77.3**	67.6	67.3
	many methods	**78.1**	73.8	76.1
Person B	-	85.2	**86.0**	85.2
	dropout	**86.9**	86.1	84.9
	weight reg.	**84.5**	84.4	83.9
	batch norm.	**86.6**	86.5	83.3
	many methods	**88.8**	87.3	86.0

4. Discussion

The following classifiers were tested during the research: dummy classifier (as baseline), basic ML algorithms and neural networks. As predicted, dummy classifier achieved very low recognition results (below 30% on average). Basic ML algorithms outperformed baseline methods but still performance was not sufficient to deploy the generated models to the target system. The best among the basic methods turned out to be the model based on the k-NN algorithm, which achieved efficiency at the level of approx. 50 | 75% (Person A | Person B). The last classifiers tested were neural networks. The best results were obtained with models based on the Multi-layer perceptron with additional regularization: 78 | 89%. It should be noted that the best results in case of the LSTM (74 | 87%) and LSTM+dense (76 | 86%) models are similar to the best results obtained by the MLP network. Therefore, on the basis of the obtained results, it is not possible to unequivocally state which of the examined types of neural network will be optimal for vocalization recognition. The tested architectures should be further researched with additional data for a larger group of people with PIMD. A promising further step seems to be examination of other variants of networks (e.g. Bidirectional LSTM). However, specific conclusions can be drawn regarding the optimal structure of the neural network. All tested neural networks performed best on the architecture based on one hidden layer with 64 neurons (for LSTM + dense: 1 LSTM layer + 1 additional dense layer). The low complexity of the network largely prevented overfitting the training data. As a result, network was capable of better generalization and thus obtained better results for the external (test) data. It can therefore be concluded that in this research problem, more extensive (deeper) networks will not be applicable. For the tested neural networks, the best method of regularization turned out to be a combination of dropout and batch norm. Both methods were set to the input and the output of the neural network layer under study. Standardization of data (batch norm.) and randomly ignoring neurons during training (dropout) also improved the ability of the network to generalize. As a result, the neural network was less prone to overfit the training data.

5. Conclusions and Future Work

The paper presents research on the use of machine learning algorithms in recognizing vocalization for people with PIMD. The best results of vocalization recognition at the level of 80% (Person A) and 90% (Person B) were obtained by the MLP neural network consisting of one layer and 64 neurons with dropout and batch normalization regularization. In order to improve the results, we plan to acquire additional vocal representations of people with PIMD, examine applicability of other neural network architectures (e.g. Bidirectional LSTM, CNN, hybrid CNN-LSTM) and various speech feature extractors (e.g. LPC, PLP, RASTA) [19]. Another promising direction is to test recognition module combining multiple machine learning methods (Ensemble Methods). In addition, we plan to include an additional pre-processing module based on empirical signal decomposition, which would allow the extraction of important components in a specific frequency bandwidth by reducing: noise, background sounds, undesirable sounds. We plan to research the usefulness of: Multivariate Variational Mode Decomposition [20], Enhanced Empirical Mode Decomposition [21], Fourier Decomposition [22], which are used in the decomposition process of non-stationary, nonlinear and highly noisy signals [23-26].

6. Acknowledgements

The research presented herewith has been conducted within the INSENSION project funded by the EU's Horizon 2020 research and innovation program under grant agreement no. 780819.

7. References

[1] M. Kosiedowski, B. Gluszak, M. Engelhardt, T. Krämer, and J. Urbanski, "Global atlas of people with profound intellectual and multiple disabilities," *Journal on Technology and Persons with Disabilities*, pp. 106–119, 2019.

[2] K. Truong and D. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, pp.144–158, 2007.

[3] T. Theodorou, I. Mporas, and N. Fakotakis, "Audio feature selection for recognition of non-linguistic vocalization sounds," in *SETN*, 2014.

[4] S. Petridis, A. Asghar, and M. Pantic, "Classifying laughter andspeech using audio-visual feature prediction," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 5254–5257, 2010.

[5] O. F. Reyes-Galaviz, C. A. Reyes-Garc´ıa, A. ´Optica, and L. E. Erro, "A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks," *9th Conference Speech and Computer*, 2004.

[6] J. Amoh and K. Odame, "Deep neural networks for identifying cough sounds," *IEEE Transactions on Biomedical Circuits and Systems,* vol. PP, pp. 1–9, 2016.

[7] F. Eyben, S. Petridis, B. Schuller, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Audiovisual classification of vocal outbursts inhuman conversation using long-short-term memory networks,", pp. 5844 – 5847, 2011.

[8] F. Eyben, S. Petridis, B. Schuller, and M. Pantic, "Audiovisual vocal outburst classification in noisy acoustic conditions," *in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 5097–5100, 2012.

[9] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing, Special Issue on Affect Analysis in Continuous Input*, 2013.

[10] A. Gujral, K. Feng, G. Mandhyan, N. Snehil, and T. Chaspari, "Leveraging transfer learning techniques for classifying infant vocalizations," in *IEEE EMBS International Conference on Biomedical Health Informatics*, 05 2019, pp. 1–4.

[11] J. Waddle, T. Thigpen, and B. Glorioso, "Efficacy of automatic vocalization recognition for anuran monitoring," *Herpetological Conservation and Biology*, vol. 4, pp. 384–388, 2009.

[12] K. Adi, M. Johnson, and T. Osiejuk, "Acoustic censusing using automatic vocalization classification and identity recognition," *The Journal of the Acoustical Society of America*, vol. 127, pp.874–83, 2010.

[13] H.-M. Chen, C.-J. H. Huang, Y.-J. Chen, C.-Y. Chen, and S.-Y.Chien, "An intelligent nocturnal animal vocalization recognition system," *International Journal of Computer and Communication Engineering*, vol. 4, pp. 39–45, 2015.

[14] Y.-J. Zhang, J.-F. Huang, N. Gong, Z.-H. Ling, and Y. Hu, "Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks," *The Journal of the Acoustical Society of America*, vol. 144, pp. 478–487, 2018.

[15] J. Stastny, M. Munk, and L. Juranek, "Automatic bird species recognition based on birds vocalization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, 2018.

[16] J. G. Selman and N. Demir, "Automatic detection foracoustic monitering of wild animals," Stanford University, Technical Report, June 2019. [Online]. Available: http://ilpubs.stanford.edu:8090/1166/

[17] H. Brugman and A. Russel, "Annotating multi-media/multi-modal resources with ELAN," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), 2004. [Online] http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf

[18] D. Namrata, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", *International Journal For Advance Research In Engineering And Technology*, Volume 1, Issue VI, 2013.

[19] K.R. Ghule and R. R. Deshmukh, "Feature extraction techniques for speech recognition: A review," *International Journal of Scientific & Engineering Research*, pp. 2229–5518, no. 5, 2015.

[20] N. Rehman and H. Aftab, "Multivariate Variational Mode Decomposition," *IEEE Transactions on Signal Processing*, vol. 67, no. 23, pp. 6039-6052, Dec. 2019.

[21] Y. Hu, F. Li, H. Li and C. Liu, "An enhanced empirical wavelet transform for noisy and non-stationary signal processing," *Digital Signal Processing*, vol. 60, pp. 220-229, Jan. 2017.

[22] P. Singh, S.D. Joshi, R.K. Patney and K. Saha, "The Fourier decomposition method for nonlinear and non-stationary time series analysis," Proc. of the Royal Society A: Math., Phys. and Eng. Sci.", vol. 473, no. 2199, art. no. 20160871, Mar. 2017.

[23] P. Cao, H. Wang and K. Zhou, "Multichannel Signal Denoising Using Multivariate Variational Mode Decomposition With Subspace Projection," *IEEE Access*, vol. 8, pp. 74039-74047, 2020.

[24] P. Kuwalek, B. Burlaga, W. Jesko, P. Konieczka, „Research on methods for detecting respiratory rate from photoplethysmographic signal," *Biomedical Signal Processing and Control*, vol. 66, art. no. 102483, Apr. 2021.

[25] P. Kuwalek, "Estimation of Parameters Associated With Individual Sources of Voltage Fluctuations," *IEEE Transactions on Power Delivery*, vol. 36, no. 1, pp. 351-361, Feb. 2021.

[26] P. Singh, A. Singhal, B. Fatimah, A. Gupta and S. D. Joshi, "AF-MNS: A Novel AM-FM Based Measure of Non-Stationarity," *IEEE Commun. Lett*., vol. 25, no. 3, pp. 990-994, Mar. 2021.