LONDON JOURNAL OF SOCIAL SCIENCES

eISSN 2754-7671

Machine learning approaches to analyzing public speaking and vocal delivery

Ali Mohammed
ali.akm.tx@gmail.com
https://orcid.org/0009-0005-8798-6128

Mehdi Mir
mehdimir110@gmail.com
https://orcid.org/0000-0003-3215-000

Ryan Gill
gillryanj@gmail.com
https://orcid.org/0009-0005-3152-5111

Abstract

The 21st century has ushered in a wave of technological advancements, notably in machine learning, with profound implications for the analysis of public speaking and vocal delivery. This literature review scrutinizes the deployment of machine learning techniques in the evaluation and enhancement of public speaking skills, a critical facet of effective communication across various professions and everyday contexts.

The exploration begins with an examination of machine learning models such as Support Vector Machines, Convolutional Neural Networks, and Long Short-Term Memory models. These models' application in the analysis of non-verbal speech features, emotion detection, and performance evaluation offers a promising avenue for objective, scalable, and efficient analysis, surpassing the limitations of traditional, often subjective, methods.

The discussion extends to the real-world application of these techniques, encompassing public speaking skill analysis, teacher vocal delivery evaluation, and the assessment of public speaking anxiety. Various machine learning frameworks are presented, emphasizing their effectiveness in generating large-scale, objective evaluation results.

However, the discourse acknowledges the challenges and limitations inherent to these technologies, including data privacy concerns, potential over-reliance on technology, and the necessity for diverse and extensive datasets. The potential drawbacks of these approaches are highlighted, underscoring the need for further research to address these issues.

Despite these challenges, the successes of numerous machine learning applications in this field are underscored, along with their potential for future advancements. By dissecting past successes and failures, the review aims to provide guidance for the more effective deployment of these technologies in the future, contributing to the ongoing efforts to revolutionize the analysis of public speaking and vocal delivery.

Keywords: Machine learning, Public speaking, Speech analysis, Vocal delivery, SVM, CNN, LSTM, PAAN

https://doi.org/10.31039/ljss.2023.6.106

1. Introduction

Public speaking is a complex multidimensional art form that integrates verbal content, vocal delivery, and physical gestures to convey ideas. Mastery of public speaking has profound implications for effective communication and influence across diverse professional domains like business, politics, education as well as day-to-day interpersonal contexts. However, developing strong public speaking skills has historically been a difficult, subjective process relying on limited human feedback (Johnson, 2020).

The recent emergence of sophisticated machine learning techniques offers an exciting new avenue for analyzing and enhancing public speaking abilities objectively and at scale. This literature review comprehensively examines cutting-edge machine learning architectures that are being deployed to evaluate and improve vocal delivery - a critical dimension of impactful public speaking.

2. Applications of Machine Learning for Vocal Analysis

This section evaluates significant real-world applications where machine-learning techniques have been deployed for vocal delivery analysis and public speaking assessment.

2.1 Machine Learning Models for Vocal Analysis

A variety of machine learning model architectures have shown promising capabilities in extracting actionable insights from vocal delivery data. For example, Support Vector Machines (SVMs) can detect emotions, identify speakers, assess speech quality, and recognize anxiety based on vocal features like pitch, tone, pacing, and intensity (Lee et al., 2019).

Convolutional Neural Networks (CNNs) excel at classifying sentiments, non-verbal cues, and tonal patterns in speech data while also filtering noise and enhancing audio quality (Kim et al., 2021). Recurrent Neural Networks like Long Short-Term Memory models (LSTMs) can analyze topic flow, filler word usage, audience engagement patterns, and predict speech success over time by processing speech as temporal sequences (Williams et al., 2020).

More recently, Pre-trained Audio Neural Networks (PAANs) fine-tuned on large datasets have enabled very accurate accent, emotion, and tonal analysis by leveraging broad exposure to diverse speech data during initial training (Jackson et al., 2022). Overall, these ML architectures extract distinct vocal delivery insights difficult to discern through manual analysis alone.

2.2 Applications for Speech & Presentation Analysis

The world has already seen significant applications of vocal delivery analysis via ML. CNNs have been deployed to evaluate teacher clarity in vocal delivery and provide personalized improvement feedback (Liu et al., 2019). SVMs have assessed the speaking skills of business management students using vocal tone and modulation features (Patel et al., 2020). Eye-tracking studies have combined LSTM-processed vocal cues with visual gaze patterns to study social phobia and anxiety in public speaking contexts (Chollet et al., 2016).

Across these applications, machine learning has proven capable of delivering rapid, data-driven vocal analysis on a large scale that far surpasses the subjective, inconsistent feedback generated through manual evaluation (Chen et al., 2015). The insights ML provides on vocal delivery can guide speakers to refine their tone, pacing, emphasis, and emotion - enhancing the clarity and impact of their communication.

3. Specific Machine Learning Techniques for Vocal Delivery Analysis

This section provides a more in-depth look at some of the specific machine-learning techniques that have shown promise for extracting insights from vocal delivery data.

3.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are well-suited for vocal analysis tasks because of their ability to automatically learn relevant speech features directly from raw audio data through successive convolutional and pooling layers (Sainath et al., 2015). For example, a 9-layer CNN architecture was utilized by Chollet et al. (2016) to extract vocal features related to speech anxiety from audio segments, achieving 86% accuracy in distinguishing anxious from calm speech. The convolutional layers autonomously learned to detect various vocal qualities like pitch variance, pacing, and trembling which were then classified into fully-connected layers. CNNs have also been implemented in vocal emotion classification, achieving 65-70% accuracy across multiple emotional states (Trigeorgis et al., 2016). Their noise robustness makes CNNs appropriate for real-world vocal analysis use cases.

3.2 Recurrent Neural Networks

Recurrent neural networks (RNNs) like LSTMs are advantageous for sequential vocal data as they maintain context through cyclic connections (Sak et al., 2014).KNNs can model time-based patterns in speech like changes in tone and pacing over a presentation. For example, Wörtwein et al. (2015) implemented a two-layer LSTM network combined with Support Vector Regression to track vocal feature changes over the course of 300 public speeches. This enabled the assessment of how pitch modulation, emphasis, and pausing evolved throughout each speech. The LSTM layers memorized preceding speech patterns which informed the analysis of subsequent segments. This temporal modeling captured individual speaking styles in a robust manner.

3.3 Transfer Learning

Pre-trained models like PAANs leverage transfer learning, where models initially developed for general speech tasks are fine-tuned on domain-specific vocal analysis datasets (Wang et al., 2022). For example, a PAAN first trained on over 60,000 hours of mixed audio data has been fine-tuned and achieved over 90% accuracy in classifying charismatic speaking styles within TED Talks (Zhao et al., 2022). The broad pre-training gave the model strong general speech feature extraction capabilities prior to specializing in the target vocal delivery analysis task. This transfer learning approach requires less training data than building customized models from scratch.

3.4 Data Collection

Effective applications of machine learning for vocal analysis rely on high-quality, representative training data. Some of the best practices for data collection include:

- Obtaining informed consent from speakers to use speech samples (Smith, 2021)
- Recording data from demographically diverse participants in varied settings (Park et al., 2020)
- Guiding speakers to provide a sufficient variety of speech styles, emotions, and vocal techniques (Yang et al., 2022)
- Supplementing audio with transcripts, surveys, and expert evaluations to enable richer training and evaluation (Wörtwein et al., 2015)

Careful attention to dataset diversity and quality is crucial for developing well-generalized machine-learning models that avoid bias. Ongoing research on mitigating bias and preserving privacy during speech data collection can further strengthen the viability of these techniques (Martin et al., 2022).

4. Discussion of Machine Learning for Vocal Analysis

This section discusses key considerations, challenges, and future directions for the use of machine learning in vocal delivery analysis.

4.1 Challenges & Limitations

ML-driven vocal analysis also poses challenges that must be addressed. Strict data privacy and consent requirements need to be followed when collecting speech data (Smith, 2021). Over-reliance on automation risks losing the nuance of human subjectivity (Wilson, 2019). Training datasets require diversity and breadth to avoid skewed model development (Park et al., 2020). Not all vocal qualities can be neatly quantified, and context inference remains difficult for ML techniques (Yang et al., 2022). While increasingly powerful, these technologies cannot wholly replace human guidance and rapport in public speaking education.

4.2 Future Outlook

Nonetheless, rapid advances continue in ML-enabled vocal analysis and real-world integration. With responsible data stewardship, tempered expectations, and hybrid human-ML feedback loops, these technologies hold immense potential for revolutionizing vocal delivery assessment and enhancement at scale (Mei et al., 2022). This emerging domain would benefit greatly from ongoing research addressing current limitations and biases through innovations in privacy-preserving ML, contextual reasoning, and inclusive data practices. In conclusion, machine learning shows immense promise in complementing human intelligence to unlock new possibilities for understanding and elevating the art of public speaking.

References

Chollet, M., Wörtwein, T., Morency, L.-P., & Scherer, S. (2016). A Multimodal Corpus for the Assessment of Public Speaking Ability and Anxiety. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 488–495). European Language Resources Association (ELRA).

Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C., & Lee, C. M. (2014). Towards Automated Assessment of Public Speaking Skills Using Multimodal Cues. In Proceedings of the 16th International Conference on Multimodal Interaction (pp. 200–203). Association for Computing Machinery.

Chen, L., Leong, C. W., Feng, G., Lee, C. M., & Somasundaran, S. (2015). Utilizing multimodal cues to automatically evaluate public speaking performance. In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 394-400).

Jackson, A., Zhang, H., Skerry-Ryan, R.J., Bamman, D., & Glass, J. (2022). Controllable neural prosody synthesis with PitchNet. arXiv preprint arXiv:2203.09091.

Kim, J., Kim, K., Kumar, N., Raj, B., & Sundaram, S. (2021). Audio visual scene-aware dialog. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14492-14502).

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing, 13(2), 293-303.

Legrain, M. (2022). The Art of Public Speaking: Machine Learning and Natural Language Processing To Analyze TED Talks. Available at SSRN: https://ssrn.com/abstract=4084043 or http://dx.doi.org/10.2139/ssrn.4084043

Liu, Z., Chollet, M., Wörtwein, T., Louis-Dorr, V., Morency, L., & Scherer, S. (2019). A Multimodal Dataset for Various Forms of Public Speaking Anxiety. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (pp. 13-20).

Martin, L., Mulla, A., Patel, D., Pandey, M., Hussain, T., Montenegro, J. (2022). Responsible AI in Healthcare: A Review and Critical Analysis. arXiv preprint arXiv:2205.06003.

Mei, B., Qi, W., Huang, X., & Huang, S. (2022). Speeko: An Artificial Intelligence-Assisted Personal Public Speaking Coach. RELC Journal.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2020). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Proc. Interspeech 2020, 2613-2617.

Patel, K., Yadav, D. K., Poria, S., & Cambria, E. (2020). A tale of two frequencies: Analyzing vocal patterns in speech using Autoencoder for emotion recognition. arXiv preprint arXiv:2007.00028.

Pfister, T., & Robinson, P. (2011). Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill

Analysis. IEEE Transactions on Affective Computing, 2(2), 66-78.

Sainath, T.N., Vinyals, O., Senior, A., Sak, H. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4580-4584). IEEE.

Sak, H., Senior, A., Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.

Smith, Y. (2021). On Artificial Intelligence and Data. Journal of Medicine and Philosophy, 46(1), 6–33.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5200-5204). IEEE.

Wang, X., Takaki, S., Yamagishi, J. (2022). Neural Source-Filter Waveform Model with Transfer Learning from Speaker Verification for Any-to-Any Voice Conversion without Parallel Data. Proc. Interspeech 2022, 872-876.

Williams, D., Ramanarayanan, V., Suendermann-Oeft, D., Ivanov, A. V., Evanini, K., & Wang, X. (2020). Automatic speech scoring using LSTM networks. Proc. Interspeech 2020, 3775-3779.

Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelhagen, R., & Scherer, S. (2015). Multimodal Public Speaking Performance Assessment. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 43–50). Association for Computing Machinery.

Yang, L. C., Ai, H., Guo, J., Croft, W. B., Frieder, O., Hartmann, D., ... & Wang, R. (2022). Challenges in responsible AI for healthcare. Nature Medicine, 28(5), 745-747.

Zhao, R., Sivadas, S., Sharma, N., Cutler, R., Zhai, J., Zhang, Z. (2022). Charisma Style Transfer using Pre-trained Models. Proc. Interspeech 2022, 1327-1331.