A NOVEL AND REALISTIC REVIEW ON RECENT TRENDS IN SPEECH PROCESSING USING PROBABILISTIC MODELS

Rubina
Senior Grade Lecturer,

Sujatha P.
Senior Grade Lecturer

Department of Computer Science Engg,
Government Polytechnic, Bagalkot, Karnataka

Abstract
Speech processing has witnessed significant advancements in recent years, with probabilistic models playing a pivotal role in enhancing various tasks, including speech recognition, synthesis, and understanding. This paper presents a comprehensive review of recent trends in speech processing, focusing on the application of probabilistic models. We explore the theoretical foundations, methodologies, and practical implementations of probabilistic models in speech processing tasks. Additionally, we discuss the challenges, emerging trends, and future directions in this rapidly evolving field.

1.Introduction

Speech processing, the interdisciplinary field concerned with the analysis, synthesis, and understanding of human speech, has experienced remarkable progress owing to advancements in machine learning and probabilistic modelling techniques. Probabilistic models, which enable the representation of uncertainty and variability in speech data, have emerged as powerful tools for addressing various challenges in speech processing tasks.

2. Theoretical Foundations of Probabilistic Models in Speech Processing:

This section provides an overview of the theoretical underpinnings of probabilistic models in speech processing. We discuss fundamental concepts such as hidden Markov models (HMMs), Gaussian mixture models (GMMs), probabilistic graphical models, and deep probabilistic models. We delve into how these models capture the statistical properties of speech signals and enable robust representation and inference.

3. Methodologies in Speech Processing Using Probabilistic Models:

In this section, we review methodologies employed in speech processing tasks leveraging probabilistic models. We explore techniques for speech recognition, where HMMs and deep neural networks (DNNs) have been integrated to improve accuracy and robustness. Additionally, we discuss speech synthesis methods based on probabilistic models, including statistical parametric synthesis and waveform synthesis. Moreover, we examine probabilistic approaches to speech understanding tasks such as speaker diarization, emotion recognition, and language modelling.

What is a speech model? It may be many things depending on the application, covering anything from tongue and jaw movements to the structure of language. Here, a speech model is a representation of the range of possible articulations of a set of predefined words or sentences. While written language is deterministic, in the sense that the spelling of a word (usually) does not vary between authors, speech is intrinsically variable. No two articulations of a particular word or sentence will ever be exactly alike, even when uttered by the same speaker.

To understand speech, a listener must disregard certain variations in spectrum, prosody and timing; something which has been proven very difficult to do automatically. The listener should also keep in mind which variations occur frequently and which are less common. This motivates the use of a probabilistic model, that expresses the probability of a given articulation, defined as PX|θ(x|θ), where x is a signal representation of the utterance and θ is the articulation model for the utterance. A big advantage of using a probabilistic speech model is that it can be used both for recognition (using e.g. MAP classification) and synthesis (by drawing samples from the distribution). This is an essential trait for the presented framework, as it enables optimal classification according to the MAP rule.

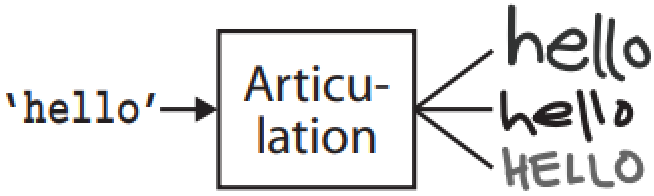


Figure 1: Due to inexact muscle control and changing

circumstances, articulation of a fixed word will always include variations in pronunciation. The situation is analogous to the relation between typed text and handwriting.

Hidden Markov models

The most commonly used probabilistic speech model is the hidden Markov model (HMM). It was originally proposed by Baum and Petrie as a way of modelling Markov sequences with noisy observations. It is now widely used for automatic speech recognition , and to a lesser extent for speech synthesis. One of the main advantages is that HMMs can be trained efficiently using the expectation maximization (EM) algorithm .

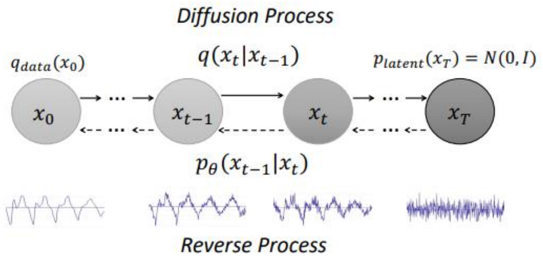
Many variations of the standard HMM formulation have been proposed in order to improve their speech modelling capabilities. One limitation in standard HMMs is the distribution of time spent in each state, which follows a geometric distribution. Several researchers have proposed to include explicit duration modelling, and although it generally complicates the training procedure, it can improve the perceptual quality of synthesized speechand may also improve automatic speech recognition performance . Another possible shortcoming is the large number of parameters required to represent a word accurately; in the standard formulation, each state has its own “output” probability distribution.

Thus, a large set of training data is required to avoid overfitting the model the data. If little training data is available, even as little as one utterance of each word, a useful modification is the tied-mixture HMM. There, a single output distribution is trained on the entire data set (most commonly a mixture model, where each component represents a phoneme or speech sound). Then, the individual word HMMs are only allowed to control the mixture weights of this distribution as their output. It should be noted that the HMM formulation of the probability distribution of speech is necessarily approximate; real speech is not generated by a first order Markov chain, but rather by an extremely complex neural and motor process.

Although attempts have been made at creating such “complete” models , they are not yet advanced enough to perform speech synthesis or speech recognition near the state of the art level.

Diffusion Probabilistic Models

This section introduces the diffusion and the reverse procedures of the diffusion probabilistic model. A detailed mathematical proof of the model’s ELBO can be found in [30], and we only discuss the diffusion and reverse processes with their algorithm in this section



Diffusion Process Reverse Process Fig. 2.

The diffusion process (solid arrows) and reverse processes (dashed arrows) of the diffusion probabilistic model.

4. Practical Implementations and Applications:

Here, we present real-world applications of probabilistic models in speech processing. We highlight successful implementations in automated transcription systems, virtual assistants, speech-enabled devices, and communication aids for individuals with speech impairments. Furthermore, we discuss the integration of probabilistic models with emerging technologies such as natural language processing, machine translation, and multimodal interaction.

5. Challenges and Future Directions:

Despite significant progress, several challenges persist in the application of probabilistic models to speech processing. These include handling variability in speech signals, addressing data scarcity, enhancing robustness to noise and environmental conditions, and improving model interpretability and generalization. We discuss ongoing research efforts and propose future directions aimed at addressing these challenges, such as leveraging deep generative models, integrating domain knowledge, and exploring multi-task learning approaches.

6. Conclusion:

In conclusion, this paper provides a comprehensive review of recent trends in speech processing using probabilistic models. We have discussed the theoretical foundations, methodologies, practical implementations, challenges, and future directions in this rapidly evolving field. Probabilistic models continue to play a crucial role in advancing the state-of-the-art in speech processing, offering promising avenues for further research and innovation.

References:

[1] Articulation index predictions for hearing-impaired listeners with and without cochlear dead regions. Journal of the Acoustical Society of America, 111(6):2545–2548, 2002.

[2] J. B. Allen. Harvey Fletcher's role in the creation of communication acoustics. Journal of the Acoustic Society of America, 99(4 pt 1):1825– 1839, 1996.

[3] U. Andersson. Cognitive deafness: The deterioration of phonological representations in adults with an acquired severe hearing loss and its implications for speech understanding. Ph d thesis, Link¨oping University - Disability Research, Link¨oping, 2001.

[4] ANSI-S3.5. American national standard methods for the calculation of the speech intelligibility index. American National Standards Institute, New York, 1997.

[5] S. Arlinger. Negative consequences of uncorrected hearing loss–a review. International Journal of Audiology, 42 Suppl 2:2S17–20, 2003.

[6] R. J. Baker and S. Rosen. Auditory filter nonlinearity across frequency using simultaneous notched-noise masking. Journal of the Acoustical Society of America, 119(1):454–462, 2006.

[7] L. E. Baum and J. A. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Statist., 37:1554– 1563, 1966.

[8] J. R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter modeling for speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 38(12):2033–2045, 1990.

[9] J. J. Briaire and J. H. M. Frijns. Field patterns in a 3D tapered spiral model of the electrically stimulated cochlea. Hearing Research, 148:18– 30, 2000.

[10] S. Buus, E. Schorer, M. Florentine, and E. Zwicker. Decision rules in the detection of simple and complex tones. Journal of the Acoustical Society of America, 80(6):1646–1657, 1986.

[11] C. Cherry. On Human Communication: A Review, a Survey, and a Criticism. Wiley, 1957.

[12] T. Y. C. Ching, H. Dillon, and D. Byrne. Speech recognition of hearingimpaired listeners: predictions from audibility and the limited role of high-frequency amplification. Journal of the Acoustical Society of America, 103(2):1128–1140, 1998..]

[13] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, 2014.

[14] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 708–712.

[15] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[16] Daniel Michelsanti and Zheng-Hua Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," arXiv preprint arXiv:1709.01703, 2017.

[17] Ying-Hui Lai, Fei Chen, Syu-Siang Wang, Xugang Lu, Yu Tsao, and Chin-Hui Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," IEEE Transactions on Biomedical Engineering, vol. 64, no. 7, pp. 1568–1578, 2016