A Temporal Extension of Latent Dirichlet Allocation for Unsupervised Acoustic Unit Discovery

Werner van der Merwe, Herman Kamper, Johan du Preez
MediaLab, E&E Engineering, Stellenbosch University, South Africa

Unsupervised acoustic unit discovery

- Annotating speech is often not possible for many low-resource languages.
- Unsupervised acoustic unit discovery (AUD) aims to solve this by finding a set of phone-like units from speech that resembles the phonetic inventory of a language.

Vector-quantised neural networks

- Recently vector-quantised (VQ) neural networks such as the variational autoencoder (VQ-VAE) and contrastive predictive coding (VQ-CPC) have performed well in unsupervised AUD.
- Despite good performance, the number of VQ codes in these models (512 for the VQ model in this paper) is often far more than the number of true phone units used in a language (in the order of 50).
- To more closely match the phonetic inventory of a language, the larger set of VQ codes therefore need to be mapped to a smaller set.

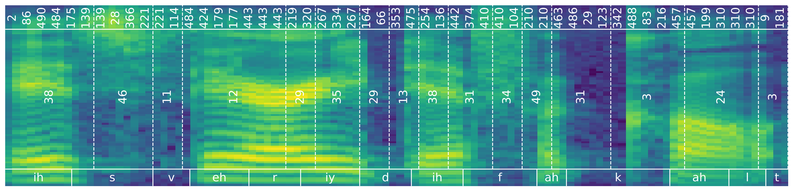

Figure: This figure gives a spectrogram of the spoken utterance, "Is very difficult". The labels on the figure from top to bottom are from a VQ-VAE, Markov chain latent Dirichlet allocation (MCLDA) and a professional transcriber.

Latent Dirichlet allocation

- Latent Dirichlet allocation (LDA) is widely used for unsupervised topic modelling on sets of documents.
- The LDA model is a parametric bag-of-VQ-codes model and considers no temporal information.

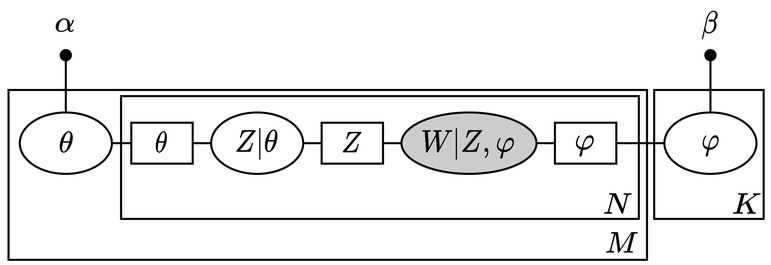

Figure: The cluster graph of latent Dirichlet allocation model.

Markov chain latent Dirichlet allocation

- We extend LDA to include a Markov chain between adjacent phone units to model transitions.
- With this model we introduce the scaler parameter a to increase the probability of repeating phone-like units.

$$p(\boldsymbol{Z}_n, \boldsymbol{Z}_{n-1}) = \begin{cases} \frac{a}{K^2+K(a-1)}, & Z_n = Z_{n-1} \\ \frac{1}{K^2+K(a-1)}, & Z_n \neq Z_{n-1} \end{cases}$$

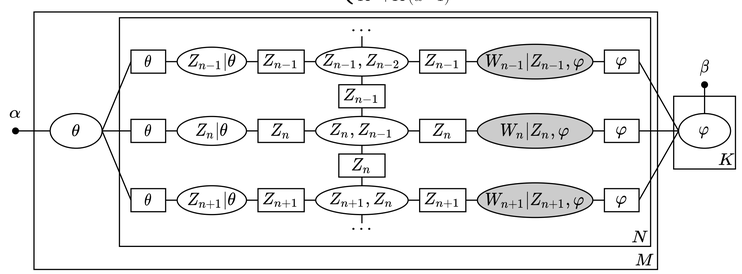

Figure: The cluster graph of Markov chain latent Dirichlet allocation model.

Experimental setup

- A VQ-VAE model is used to extract utterances with 512 VQ codes from the Buckeye speech corpus.
- The Markov chain LDA model uses these VQ codes to discover 50 latent phone-like units.
- We compare the base LDA and expanded Markov chain LDA to an established CPC-Big + K-means model that clusters CPC representations directly into 50 units.

Metrics

- We compare the output of models to the ground truth transcriptions in terms of phone segmentation and cluster quality.

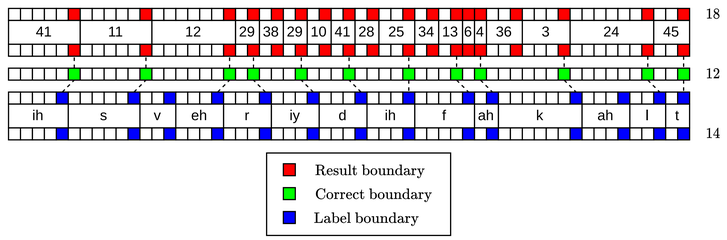

Figure: The following figure illustrates how to determine if a phone boundary is correct.

Results: Phone Segmentation

Table: Phone segmentation results (%) on Buckeye speech test. 100% in all metrics would be perfect segmentation. Negative R-value

Model	*Precision*	*Recall*	*F-score*	*R-value*
Speaker dependent				
CPC-Big + km-50	35.3	94.0	51.3	−44.4
VQ-VAE	32.1	**97.6**	48.1	−80.7
LDA	36.0	94.9	52.00	−45.0
Markov chain LDA	**55.1**	72.4	**62.2**	**55.6**
Speaker independent				
CPC-Big + km-50	35.5	93.9	51.6	−42.5
VQ-VAE	32.0	**97.7**	48.2	−76.2
LDA	36.8	93.3	52.8	−33.4
Markov chain LDA	**55.4**	66.4	**60.4**	**61.6**

Results: Cluster quality

Table: Clustering quality results (%) on Buckeye speech test set. The desired purity and mutual information is 100% while a lower singleton % is better.

Model	*Cluster Purity*	*Singletons*	*Mutual Information*
Speaker dependent			
CPC-Big + km-50	33.1	9.3	**39.6**
VQ-VAE	**34.0**	26.7	38.4
LDA	24.1	20.2	26.6
Markov chain LDA	29.1	**4.5**	30.6
Speaker independent			
CPC-Big + km-50	**32.3**	9.3	**37.8**
VQ-VAE	31.8	27.0	34.3
LDA	21.5	18.9	22.0
Markov chain LDA	27.3	**4.0**	24.7

Conclusions

- We have shown that the inclusion of temporal information in an LDA model increases the AUD phone segmentation and mutual information.
- The Markov chain LDA achieved better phone segmentation results than the base LDA and the CPC-Big + K-means method.
- The mutual information from our Markov chain LDA model was lower than that of the CPC-Big + K-means method we compared against, but achieved better phone segmentation results.
- The introduction of more informative priors could result in an increased mutual information.

Poster presented at Interspeech 2022. [Incheon Korea, September 2022]