Doctoral Dissertation

DPGMM-RNN Hybrid Model: Towards Universal Acoustic Modeling to ASR at Different Supervised Levels

Bin Wu

Program of Information Science and Engineering
Graduate School of Science and Technology
Nara Institute of Science and Technology

Supervisor: Professor Satoshi Nakamura
Augmented Human Communication Laboratory (Division of Information Science)

Submitted on July 20, 2021

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering

Bin Wu

Thesis Committee:

Supervisor	Satoshi Nakamura
	(Professor, Division of Information Science)
	Taro Watanabe
	(Professor, Division of Information Science)
	Jinsong Zhang
	(Professor, Graduate School of Information Science, BLCU university)
	Sakriani Sakti
	(Associate Professor, School of Information Science, JAIST)

DPGMM-RNN Hybrid Model: Towards Universal Acoustic Modeling to ASR at Different Supervised Levels*

Bin Wu

Abstract

The independent development of methods for unsupervised and supervised learning induces the different treatments to the unsupervised phoneme discovery and the supervised speech recognition; the two tasks both need acoustic modeling to find patterns that form the perceptual units such as phonemes and words; the only difference is at different supervised levels. So it is reasonable to regard the unsupervised phoneme discovery as the unsupervised ASR (that finds units from speech without text). We propose to use universal acoustic modeling (instead of separated ones) of supervised and unsupervised ASR for the whole process from acoustic waveform to speech units.

The study aims to construct universal acoustic modeling for speech recognition at different supervised levels. Specifically, the work proposes the hybrid model, which combines the Dirichlet process Gaussian mixture model and recurrent neural network (DPGMM-RNN). Furthermore, the proposed approach is utilized (1) to improve phoneme categorization by relieving the fragmentation problem; (2) to extract perceptual features to improve ASR performance.

Keywords:

Dirichlet process Gaussian mixture model (DPGMM), recurrent neural network (RNN), unsupervised phoneme discovery, perceptual feature extraction, ASR

*Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, July 20, 2021.

Acknowledgements

I thank the MEXT scholarship for supporting me to live in Japan, a place with appealing culture and environments. I thank Professor Satoshi Nakamura, Associate Professor Sakriani Sakti, Professor Taro Watanabe, and Professor Jingsong Zhang for providing freedom and guiding me into the interesting research world. I thank Naist for providing me with a quiet place to concentrate on practicing reasoning and exploring studies. I thank AHC lab for providing a free environment and thank the interesting discussions with students and professors for enabling me to learn to systemically inquiry to my research problems and closely examine different ideas, to source the knowledge from centuries ago to access the beautiful and thoughtful words full of wisdom, and to rebuild my whole life thoughts and research values in my heart. I thank all the authors in the papers and articles that enlight me to know and drive me to believe I should also leave something in my life that can interest or inspire others, just as the various happy moments when I read research works that feed me with the knowledge and inspire me to create new knowledge from my own empirical explorations. I thank lab members and friends for sharing interesting experiences and ideas to change and develop myself. I thank Ms. Manami Matsuda, my Japanese labmates, and all my friends for helping me to overcome problems in my life. I thank my family for always patiently accompanying and supporting me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Background

1.1.1 An Universal Model for Spoken Language Learning

The ASR aims to find speech patterns that match such perceptual units as phonemes and words. These units serve as acoustic elements and semantic encoding that are highly developed and optimized in evolution to fit the needs of conveying meaningful messages efficiently [14]. Such communicative units come into the brains of the infants, which might be 'blank papers' to be written by sensory stimuli such as speech [15]. Infants can find perceptual units before reading or writing that involves textual languages [16]. When infants grow older and have more language experiences, they learn these units by speaking, reading, and writing besides hearing. These behaviors unify in the neural learning of the brain cortex by processing the impulses that come from the ears that extract meaningful units, eyes that read the text words, and motor-controlled muscles that speak out or write down the words [17]. In such a view, the brain cortex stores such perceptual patterns as traces formed by electrical impulses from sensory organs. The neural encoding of the traces engraved by the sensory experiences serves as the physical basis of the spoken language learning. The neural learning processes leave the traces to encode with the past knowledge and retrieve the information for speech communication [18]. All these processes happen in one brain.

Inspired by the neural studies, this thesis considers the potential Automatic

Speech Recognition (ASR) model that learns its parameters from speech corpora to mimic the development of the same area within the temporal lobe [17] of the brain to store, update, and retrieve information from spoken language learning experiences starting from an infant of a blank state to adult of full development. The ultimate goal of the thesis is to construct universal acoustic modeling for ASR. Such acoustic modeling involves the whole process from acoustic waveform to speech units to realize the brain functionality [15] that finds perceptual units in speech without text (from infants to 4-year-olds), in a limited amount of speech with text (from 4-year-olds to children), and in a large amount of speech and text (from children to adults).

In this thesis, we propose the DPGMM-RNN hybrid model for the universal acoustic modeling. This proposal is a preliminary work to achieve universal methods on both unsupervised and supervised ASR tasks. Several related works discuss extensions about universal acoustic modeling. The previous work [19] combines data from different domains using a mixture model to solve the domain mismatching problem as an acoustic model universal to different domains. The previous work [20] uses a universal set of phones instead of phonemes special to a language to build an acoustic model universal to languages. There are also several works about acoustic models universal to different applications such as the acoustic sounds of music [21] and electromagnetic acoustic noise of cars [22]. The proposed method and features can be extended to such related tasks as applications universal to different domains [19], the discovery of phoneme set across languages [20], and application extension of different types of sounds such as music [21] and electromagnetic acoustic noise [22].

To explore such universal acoustic modeling for unsupervised phoneme discovery of infants and supervised speech recognition of adults, we first review the existing models in the Section 1.1.2 and Section 1.1.3, then discuss the limitation of the existing models in the Section 1.1.4, and finally state the contribution of this thesis to construct the universal acoustic models.

1.1.2 Automatic Speech Recognition

The ASR mimics the human auditory ability to decode the perceptual units from the speech signals. It is a fundamental spoken language technology that makes

our daily life convenient. We apply the ASR technology to build dialogue systems to help people with impairments, translation systems to help break the language barriers, and automatic spoken typing systems to help second language learners with the difficulty of written languages.

Early ASR methods started with the spectrum template matching and evolved to be HMM-GMM model that combines the temporal modeling of Hidden Markov Model (HMM) and the spectrum modeling of Gaussian Mixture Model (GMM) [23], which had been the standard method. In 2011, the HMM-DNN hybrid model that uses the DNN to relearn the temporal-spectrum connections and the acoustical distribution of speech segments achieved a breakthrough in Large Vocabulary Continuous Speech Recognition (LVCSR) tasks [24]. To replace HMM model with RNN for temporal modeling, Connectionist Temporal Classification (CTC) [25] and attention models [26, 8] have gained popularity. The recent success of transformer [27] in natural language processing also attracts research interest for ASR communities.

1.1.3 Unsupervised Phoneme Discovery

The infants can find the perceptual units from speech without text mainly within the first year [16] that is long before they learn reading and writing. Such unsupervised ability contrasts the mainstream success in the ASR with the neural network technology [8, 24] by training a large amount of paired data of speech and text. Inspired by infant studies, the research of finding phoneme-like units from audio without text (the unsupervised phoneme discovery) has attracted the interest of researchers. Such technology can help fieldworks of linguists in documenting languages without written form (which needs expert linguistic knowledge and professional auditory training) [6] and help automatic linguistic unit annotation of low-resource languages to build ASR systems (which costs money and time) [7]. The model simulation of the unsupervised phoneme discovery also helps illustrate the infant learning process [28].

In recent years, researchers are holding several Zerospeech challenges [29, 30, 3] to provide the same datasets and measurements including the ABX discrimination test [31] for fair comparision of different models in ability to discriminate the phonemes. The models in Zerospeech includes neural representation learning

by autoencoder [32, 33, 34], neural discriminative training by ABnet [35], neural discretized learning by VQ-VAE [36], traditional clustering by GMM [36] and K-means [37, 36], and nonparametric clustering by the Dirichlet Process Gaussian Mixture Model (DPGMM) trained with Gibbs sampling [38] or variational inference [39, 40].

Zerospeech Challenges compared various unsupervised models. The DPGMM feature achieved state-of-art performances from Zerospeech 2015 to Zerospeech 2019. The VQVAE and VQCPC features achieved a competitive performance and state-of-art performance in Zerospeech 2019 and 2020 respectively. Next, I will point out the problems of such VAE-based methods as the VQVAE and VQCPC, and compare them with our proposed DPGMM-based methods.

Both VQVAE and VQCPC originate from the autoencoder. The autoencoder extracts the hidden representation from a middle layer of a neural network that maps each feature to itself. The extracted over-rich hidden representation can easily copy unimportant details of the original data; the autoencoder representation loses the generalization power.

The variational autoencoder (VAE) partially solves the problem by constraining the over-rich representation. The VAE model appends the reconstruction loss with a regularization term that constrains over-rich encoded representation to be simple, distributed as close to a Gaussian or uniform prior distribution as possible.

The VAE suffers from variational approximation in its maximum likelihood estimation. The variational approximation includes two approximations: 1) variational assumption and 2) neural network approximation. Firstly, VAE uses the variational assumption. The VAE aims to minimize the KL distance between the variational distribution and the posterior distribution for each data point. The VAE usually assumes the variational distribution to be a simple mathematically-convenient Gaussian. Such an assumption make KL distance never become zero to cause suboptimal solutions to the maximum likelihood estimator. Secondly, under the variational assumption, the VAE uses neural network approximation. Encoder infers a mean and variance that minimizes KL distance. Such a huge search space (of Gaussian functions for every single data point) makes the neural network approximate, but never precisely reach, the exact analytic solutions (of

KL minimization for most data points in practice). The neural network sacrifices the accuracy (to solve the KL optimization problem for each single data point) to boost the efficiency (to approximate the KL optimization problem for many data points simultaneously).

The VQVAE constrains the VAE to encode tons of continuous feature vectors into a limited number of continuous features. Such quantization decreases the complexity of the original input signal for downstream tasks. It also decreases the bit rate. However, error rate and bit rate tradeoff might occur that the quantization might remove important properties of the original signal. The VQCPC uses CPC discriminative training on top of VQVAE.

Both VQVAE and VQCPC are based on VAE that suffers from over-rich hidden representation and variational approximation. Compared with VAE approximation of such VAE-based methods as VQVAE and VQCPC that rarely converges to the ideal maximum likelihood estimator. The DPGMM trained with MCMC sampling has no approximation and is guaranteed to converge according to detailed balance condition of Markov chain theory. The DPGMM can be quite powerful on the Gaussian-distributed speech features.

Our proposed DPGMM-RNN model improves DPGMM. The VQCPC model improves VQVAE. We will compare the DPGMM-RNN model and VQCPC model by low-resource ASR task.

The DPGMM model is a graphical model with a few parameters, each has its causal meaning. In contrast, the VQVAE based model uses neural networks with many parameters. The meaning of these model parameters and the interpretability of the model features are still left as important open research problems in the study of unsupervised phoneme discovery.

Compared to the unsupervised VQVAE based model, our proposed DPGMM-RNN hybrid model combines the unsupervised DPGMM module with a supervised RNN module. Recently, the supervised transformer model has a great success in NLP in learning distant linguistic relations. The success in NLP suggests that, in speech processing, the transformer can explore the underlying linguistic structure. Further linguistic and semantic analysis of DPGMM hybrid models with supervised transformer and RNN modules would be challenging but meaningful, which is an open problem to the study of unsupervised phoneme discovery.

Compared to the simple and general model assumption of DPGMM, the DPGMM extension as a more complex graphical model can be effective only if the causal assumption of speech data is right. Such extension is difficult due to the complexity of speech data from the real world and the naive linguistic knowledge of humans. But once successful, it rewards us with meaningful parameters to increase our knowledge.

1.1.4 Limitation

The independent development of methods for unsupervised and supervised learning induces the different treatments to the unsupervised phoneme discovery and the supervised speech recognition.

However, both Zerospeech and ASR tasks need acoustic modeling to find patterns that form the perceptual units such as phonemes and words; the only difference is at different supervised levels. So it is reasonable to regard the Zerospeech task as the unsupervised ASR to find the units without text information. In this thesis, we will explore the methods for universal acoustic modeling to ASR at different supervised levels, including unsupervised phoneme discovery and automatic speech recognition.

1.2. Thesis Contribution

This thesis has the following contribution.

- We combined DPGMM and neural network (to construct DPGMM-RNN hybrid model) to improve unsupervised phoneme discovery.
- We used DPGMM and DPGMM-RNN hybrid model as universal methods to improve acoustic modeling of ASR at different supervised levels — including unsupervised phoneme discovery, low-resource ASR, and LVCSR (with a comparison with bottleneck features from Kaldi).
- We proposed a new direction of tackling unsupervised phoneme discovery in an alternative perception-driven approach besides the data-driven approach.

- We analyzed the relation between DPGMM based features and infant auditory perception. We used DPGMM based features as perceptual features that can model infant speech learning experiences to improve ASR performance of MFCC features as the sensational features that fail to model the influence from past experiences.

1.3. Thesis Outline

The thesis is arranged as follows.

- The second chapter introduces the DPGMM clustering algorithm and the proposed DPGMM-RNN hybrid model; it also introduces the proposed perceptual feature extraction from these models for ASR.
- The third chapter introduces the first proposal of the DPGMM-RNN hybrid model to improve unsupervised phoneme discovery.
- The fourth chapter introduces the second proposal of DPGMM and DPGMM-RNN perceptual features to improve the low-resource ASR and LVCSR.
- The final chapter concludes that the two proposals provide universal methods to improve acoustic modeling at the feature level to ASR at different supervised levels.

Chapter 2

DPGMM Based Methods

2.1. Basic DPGMM Clustering Algorithm

2.1.1 Definition of DGPMM as a Generative Model

We can view each frame of a speech feature as one sample generated by a Gaussian Mixture Model (GMM) for the following reasons. Theoretically, a GMM has the power to model any distribution, especially spherical or elliptical ones with multiple local modes; practically, the GMM is suitable to fit the spectrum feature, as done in a HMM-GMM hybrid system [41] for speech recognition.

More specifically, if our data have a sequence of speech features $X = X_1, \ldots, X_n$, then each speech feature X_i has a distribution of K Gaussians mixed with weight $\pi_1, \ldots, \pi_K$:

$$p(x_i) = \sum_{k=1}^{K} \pi_k p(x_i|\mu_k, \Sigma_k)\,. \tag{2.1}$$

When clustering, we can generate each speech feature X_i in an equivalent way to Eq. (2.1). Sample one hidden Gaussian cluster Z_i and feature X_i from that Gaussian:

$$p(x_i) = \sum_{k=1}^{K} p(Z_i = k)p(x_i|Z_i = k), \tag{2.2}$$

where $p(Z_i = k) = \pi_k$ and $p(x_i|Z_i = k) = p(x_i|\mu_k, \Sigma_k)$.

The Dirichlet Process Gaussian Mixture Model (DPGMM) is a nonparametric Bayesian [42] version of GMM. The number of clusters K is learned from data

X **nonparametrically** (the number of parameters can grow with the data size); in the **Bayesian** world, our parameters are no longer unknown constants, but random variables with certain distributions. Now we generate the parameters (the weight, the mean, and the variance for each Gaussian) from their prior distributions.

Generate weight

We generate weights $\pi_1, \ldots, \pi_k, \ldots$ by the stick-breaking process [43]. Imagine a stick with length 1. We break off the stick with lengths 0.5 and 0.5 remain, and then we break off the remainder with lengths 0.25 and 0.25 remain, and we can break off more pieces in this manner forever whose lengths are $0.5, 0.25, 0.125, 0.0625, \ldots$. Finally, we get an infinite sequence of numbers that represents the mixture weights of Gaussians $(\pi_1, \pi_2, \pi_3, \ldots = 0.5, 0.25, 0.125, \cdots)$.

If every time we randomly break the previous remainder (length: r_{k-1}) into a k-th piece (length: $V_k \cdot r_{k-1}$) and the next remainder (length: $(1 - V_k) \cdot r_{k-1}$) with random proportion $V_k \sim Beta(1, \alpha)$, then we get an infinite sequence of random variables known as the stick-breaking process. We generate weight $\pi_1, ..., \pi_k, ...$ by the stick-breaking process (Algorithm 1).

Algorithm 1 Stick-breaking Process for Generating Weights

```
Draw V_1, V_2, ··· ~ Beta(1, α)
Let π_1 = V_1
Let r_1 = 1 − V_1
for k = 2, 3, ... do
    Let π_k = V_k · r_{k−1}
    Let r_k = (1 − V_k) · r_{k−1}
end for
```

When applying the DPGMM clustering for unsupervised phoneme discovery, parameters α are set according to our prior knowledge before we see the data:

- If we only believe a few phonemes are extremely frequently used in the speech, we should set parameter α to small (according to Algorithm 1. If α is small, then V tends to be large; so after several breaks of the stick with

very long pieces, the remainder will be very small for creating new Gaussian clusters.

Generate mean and variance

We generate mean and variance $(\mu_1, \Sigma_1), \ldots, (\mu_k, \Sigma_k), \ldots$ by sampling from the normal-inverse-Wishart distribution $\mathcal{NIW}(\mu_0, \lambda, \Sigma_0, \nu)$ [44] with the density function:

$$\begin{aligned} p(\mu, \Sigma) &= p(\Sigma) \cdot p(\mu|\Sigma) \\ &= \mathcal{W}^{-1}(\Sigma|\Sigma_0, \nu) \cdot \mathcal{N}(\mu|\mu_0, \frac{1}{\lambda}\Sigma)\,, \end{aligned} \tag{2.3}$$

where μ_0 and Σ_0 are the prior beliefs of the mean and the variance and λ and ν are our strengths of the beliefs of the mean and the variance. $\mathcal{W}^{-1}$ is the inverse Wishart distribution. By its definition, we immediately get the expectation of the variance (of inverse-Wishart distribution) and the variance of the mean (of normal-inverse-Wishart distribution):

$$\mathrm{E}(\Sigma) = \frac{1}{\nu - d - 1}\Sigma_0\,, \tag{2.4}$$

$$\mathrm{Var}(\mu|\Sigma) = \frac{1}{\lambda}\Sigma\,, \tag{2.5}$$

where d is the dimension of the feature ($\nu > d + 1$).

When applying DPGMM clustering to unsupervised phoneme discovery, the parameters $(\mu_0, \lambda, \Sigma_0, \nu)$ are set by our prior knowledge before we see the data:

- Prior beliefs of the mean (μ_0) and variance (Σ_0) can be approximated by the sample mean and variance of speech features from the data or by some empirical knowledge.
- If we believe the phonemes sound very different from each other, we should set the belief-strength of mean (λ) to a small value (Eq. (2.5)).
- If we believe each phoneme has high pronunciation variation, we should set the belief-strength of variance (ν) to a small value (Eq. (2.4)).

Generate data from parameters

We generated the parameters of the weight, the mean, and the variance for each Gaussian by sampling from the stick-breaking process (Algorithm 1) and the normal-inverse-Wishart distribution (equation (2.3)). Now we sample our data as we do in the GMM model (Eq. (2.2)): sample one hidden Gaussian cluster Z_i according to the weights and sample feature X_i from that Gaussian.

The DPGMM($\alpha, \text{NIW}(\mu_0, \lambda, \Sigma_0, \nu)$) can be defined by its data generation process or equivalently be represented by a Bayesian network (Fig. 2.1).

Summary: the DPGMM definition as a graphical model

We give a summary of the definition of DPGMM: we treat DPGMM as an infinite GMM with density $p(x_i) = \sum_{k=1}^{\infty} \pi_k p(x_i|\mu_k, \Sigma_k)$ (alternatively with an auxiliary hidden variable, $p(x_i) = \sum_{k=1}^{\infty} p(Z_i = k)p(x_i|Z_i = k)$). DPGMM is defined as a graphical model (Fig. 2.1) with the following generative process.

- It generates mixture weights $\{\pi_k\}_{k=1}^{\infty}$ from a stick-breaking process [43] with concentration parameter α;
- it generates means and variances $\{\mu_k, \Sigma_k\}_{k=1}^{\infty}$ from normal-inverse-Wishart (NIW) distribution with a belief of mean μ_0, a belief of variance Σ_0, a belief-strength of mean λ, and a belief-strength of variance ν; the NIW distribution has the parameter $\beta = (\mu_0, \lambda, \Sigma_0, \nu)$;
- it generates a hidden variable $Z_i = k$ by mixture weights $\{\pi_k\}_{k=1}^{\infty}$; the hidden variable indicates that the i-th data point is generated by k-th Gaussian;
- it generates each data point X_i by the Gaussian with mean μ_k and variance Σ_k indicated by the hidden variable $Z_i = k$.

We summarize this generative procedure for the graphical model of DPGMM and describe the dependency relations among the random variables of the model in Fig. 2.1.

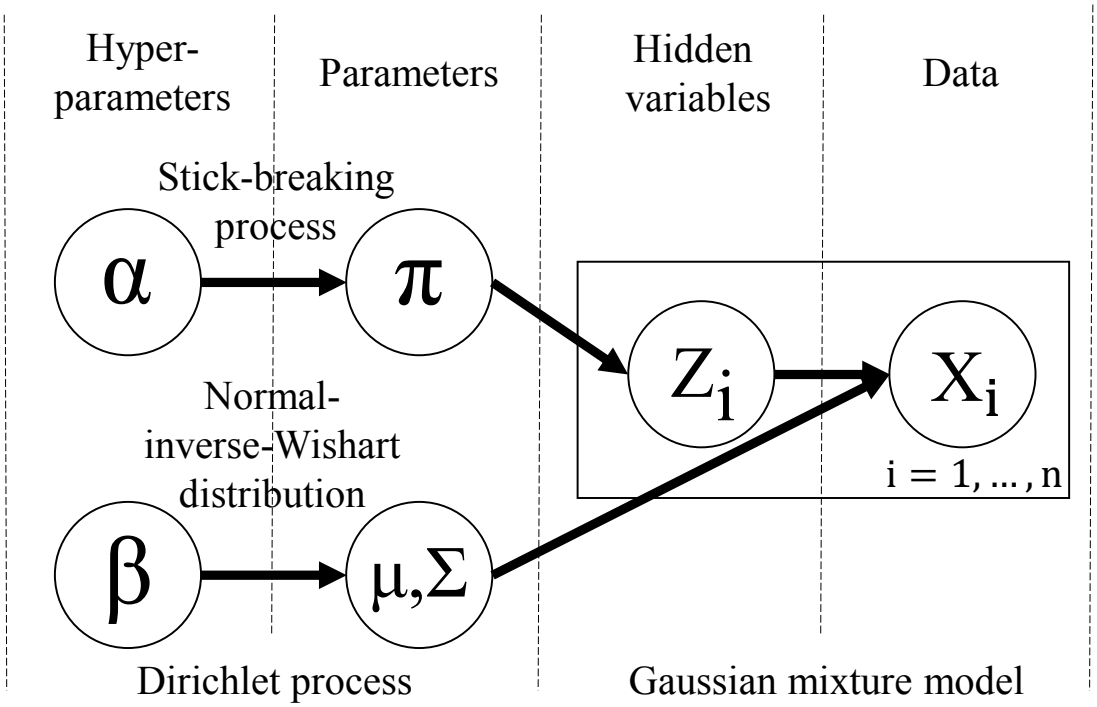


Figure 2.1: The graphical model of Dirichlet Process Gaussian Mixture Model (DPGMM) generates parameters of weights ($\pi = \pi_1, \ldots, \pi_k, \ldots$), means, and variances ($(\mu, \Sigma) = (\mu_1, \Sigma_1), \ldots, (\mu_k, \Sigma_k), \ldots$) of Gaussians from stick-breaking process (with concentration parameter α) and normal-inverse-Wishart distribution (with parameter $\beta = (\mu_0, \lambda, \Sigma_0, \nu)$) respectively; it generates hidden indicator variable $Z_i = k$ according to weights; it generates each frame of speech feature X_i (of data $X = X_1, \ldots, X_n$) by one Gaussian with mean μ_k and variance Σ_k indicated by hidden variable $Z_i = k$. The box, with (Z_i, X_i) inside, is all n data points (features) with their indicator hidden variables $((Z_1, X_1), \ldots, (Z_i, X_i), \ldots, (Z_n, X_n))$.

2.1.2 Posterior Inference of DPGMM by Gibbs Sampling

Algorithm 2 Gibbs sampling for DPGMM (Fig. 2.1) given hyperparameters α and β and observed data x

Randomly initialize cluster indicator $z = z_1, ..., z_n$
for Iteration $iter = 1, 2, \ldots$ **do**
  Sample $\pi' \sim p(\pi|z, \alpha)$ by Eq. (2.9),
    $\pi_1, \cdots, \pi_K, \pi^*_{K+1}|z, \alpha \sim \text{Dir}(n_1, n_2, \cdots, n_K, \alpha)$
  Sample $\mu', \Sigma' \sim p(\mu, \Sigma|z, \beta, x)$ by Eq. (2.10),
    $\mu_k, \Sigma_k|z, \beta, x \sim \text{NIW}(\mu_0^{(k)}, \lambda^{(k)}, \Sigma_0^{(k)}, \nu^{(k)})$
  Sample $z'_i \sim p(z_i|\pi', \mu', \Sigma', x_i)$ by Eq. (2.12),
    $z_i|\pi, \mu, \Sigma, x_i \sim \pi_k p(x_i|\mu_k, \Sigma_k)/p(x_i)$
  Update $z = (z'_1, ..., z'_n)$.
end for

After the DPGMM sees the data, we can update it by Gibbs sampling [45]. Assume that we already have a set of the hyperparameters of α and β ($\beta = (\mu_0, \lambda, \Sigma_0, \nu)$) and want to update the hidden indicator variables ($Z = Z_1, \ldots, Z_n$) and the parameters of the weights ($\pi = \pi_1, \ldots, \pi_k, \ldots$), the means and the variances ($(\mu, \Sigma) = (\mu_1, \Sigma_1), \ldots, (\mu_k, \Sigma_k), \ldots$) for the Gaussians based on the observed data ($x = x_1, \ldots, x_n$).

The Gibbs sampling of DPGMM iteratively samples the hidden variables or the parameters conditioned on all other variables until it converges. We can simplify the sampling process by conditioning on the 'surrounding' variables (the Markov blankets [46]) instead of 'all other' variables, based on the Bayesian network representation (Fig. 2.1) of DPGMM:

$$p(\pi|z, \mu, \Sigma, \alpha, \beta, x) = p(\pi|z, \alpha) \tag{2.6}$$

$$p(\mu, \Sigma|z, \pi, \alpha, \beta, x) = p(\mu, \Sigma|z, \beta, x) \tag{2.7}$$

$$p(z_i|\pi, \mu, \Sigma, \alpha, \beta, x_i) = p(z_i|\pi, \mu, \Sigma, x_i). \tag{2.8}$$

More specifically, the Gibbs sampling of DPGMM, similar to the EM algorithm, iterates over two steps until it converges.

Fix the hidden variables Z=z, estimate the parameters

First, we update the weights by sampling from a Dirichlet distribution:

$$\pi_1, \cdots, \pi_K, \pi^*_{K+1} | z, \alpha \sim \mathrm{Dir}(n_1, n_2, \cdots, n_K, \alpha)\,, \tag{2.9}$$

where K is the number of the clusters of the currently observed data, $\pi^*_{K+1} = \sum_{k=K+1}^{\infty} \pi_k$ is the sum of the weights for the future possible clusters, and $n_k = \sum_{i=1}^{n} \delta(z_i = k)$ is the number of data points in cluster k, which is counted by given hidden indicator variables $z = z_1, \ldots, z_n$.

As shown in Eq. (2.9),

- the more data (n_k) we see in Gaussian cluster k, the more weight π_k we assign to that cluster (because of expectation $\mathrm{E}(\pi_k) = \frac{n_k}{\alpha + \sum_k^K n_k}$);
- we always leave some chance of creating new clusters for future observed data by α; the smaller α is, the less the tendency to create a new cluster (because of expectation $\mathrm{E}(\pi^*_{K+1}) = \frac{\alpha}{\alpha + \sum_k^K n_k}$).

Second, we update the mean and the variance for each Gaussian cluster k by sampling a normal-inverse-Wishart distribution after seeing data x:

$$\mu_k, \Sigma_k | z, \beta, x \sim \mathrm{NIW}(\mu_0^{(k)}, \lambda^{(k)}, \Sigma_0^{(k)}, \nu^{(k)})\,, \tag{2.10}$$

where

$$\begin{aligned}
\mu_0^{(k)} &= \frac{\lambda}{\lambda + n_k} \cdot \mu_0 + \frac{n_k}{\lambda + n_k} \cdot \bar{x}_k \\
\lambda^{(k)} &= \lambda + n_k \\
\nu^{(k)} &= \nu + n_k \\
\Sigma_0^{(k)} &= \Sigma_0 + S_k + \frac{\lambda n_k}{\lambda + n_k} (\bar{x}_k - \mu_0)(\bar{x}_k - \mu_0)^T
\end{aligned} \tag{2.11}$$

with

$$\bar{x}_k = \frac{1}{n_k} \sum_{i=1, z_i = k}^{n} x_i;\ S_k = \sum_{i=1, z_i = k}^{n} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T\,.$$

As shown in Eqs. (2.10), (2.16), if we have more data points n_k in Gaussian cluster k,

- the cluster becomes more stable during sampling, the sampled centers (μ_k) become closer to each other, the sampled variances (Σ_k) become smaller (see Eqs. (2.4), (2.5) when belief-strengths ($\lambda^{(k)}$, $\nu^{(k)}$) become larger);

- our posterior belief of the mean ($\mu_0^{(k)}$) of cluster k is closer to the sample mean ($\bar{x}_k$) of the cluster. The posterior belief of the variance ($\Sigma_0^{(k)}$) will be the summation of the prior belief of the variance (Σ_0), the sample variance (S_k) of the cluster, the deviation (between the data center ($\bar{x}_k$), and the prior center (μ_0)).

Fix the parameters and infer the hidden variables

We update the hidden variables by sampling the posterior distribution:

$$p(z_i = k|\pi, \mu, \Sigma, x_i) = \frac{\pi_k p(x_i|\mu_k, \Sigma_k)}{p(x_i)} \propto \pi_k p(x_i|\mu_k, \Sigma_k)\,. \tag{2.12}$$

After the Gibbs sampling (Algorithm 2) converges, we can infer the cluster for each speech feature (data point) by the posterior:

$$z_i^* = \text{argmax}_k\, p(z_i = k|\pi, \mu, \Sigma, x_i)\,, \tag{2.13}$$

where the posterior can be computed by Eq. (2.12).

2.2. Proposed Extended DPGMM Methods

2.2.1 RNN Structure with Temporal Contextual Enhancements

The RNN structure is an extension of the feedforward neural network structure emphasizing in modeling temporality. The feedforward neural network is a stack of linear layers. Each node of a linear layer is a parameter-weighted linear combination of nodes from the previous layer that passes through an activation function such as the sigmoid function (Fig. 2.2 (a)):

$$h_t \quad = \sigma(W_{xh}x_t + b_h)\,, \tag{2.14}$$

where σ is the sigmoid function, h_t is the hidden state, W_{xh} is the weighted matrix between the input x_t and the hidden state h_t, and b_h is the bias. The RNN is an infinite sequence of feedforward neural networks with one at each time step. These feedforward neural networks share identical parameters at each layer and accumulate information through time steps by temporal connections in certain layers (Fig. 2.2 (b)):

$$h_t \quad = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h)\,, \tag{2.15}$$

where W_{xh} is the weighted matrix between the input and the hidden state, W_{hh} is the weighted matrix between the previous hidden state h_{t-1} and the current hidden state h_t, and b_h is the bias. Alternatively, rather than a sequence of the parameter-shared feedforward neural networks, the RNN is a single feedforward neural network with feedback loops at certain layers. The feedback loops of RNN accumulate temporal information (Fig. 2.2 (b)). The loops also accumulate the multiplications of gradient matrice backpropagated in RNN training. Such matrice multiplication of small neural network gradients causes the problem of the vanish of the gradient. The problem of the vanish of the gradient is relieved

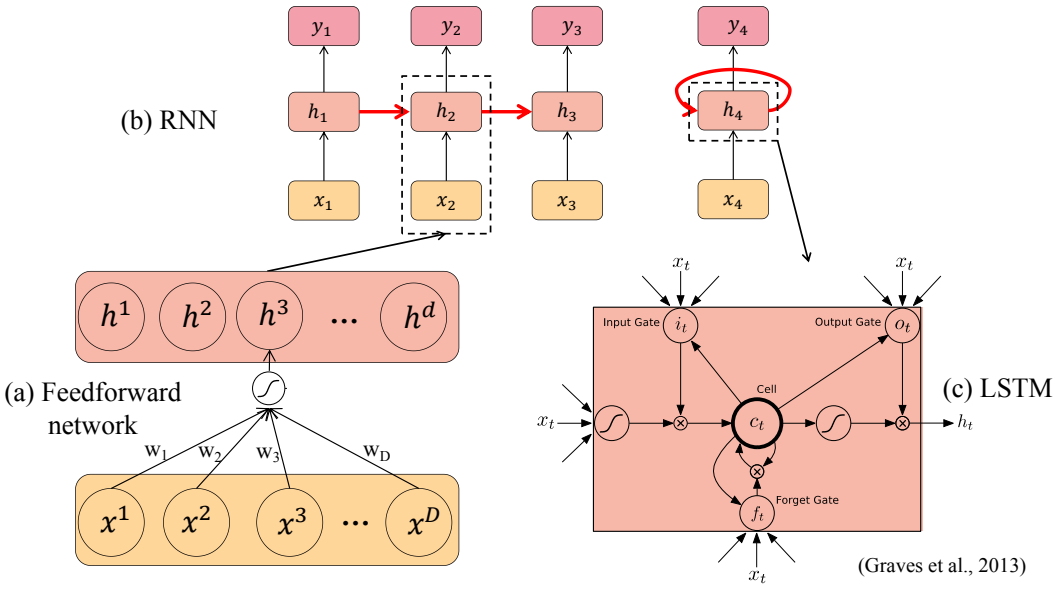


Figure 2.2: Structure of an RNN. The bottom left subfigure is (a) feedforward network that is the basic unit of RNN. A feedforward network feeds an activated weighted linear combination of dimensions of a feature to the next layer. The top subfigure is (b) RNN that is a sequence of copies of the feedforward networks with temporal information flowing through hidden units. The bottom right subfigure is (c) LSTM cell that is the extension of an RNN hidden state. An LSTM cell includes the input gate that masks out the standard RNN inputs, the forget gate that masks out the previous cell, the cell that stores a mixture from the input and the forget gates, and the output gate that masks out the values of the current hidden states. We emphasize the temporal information flow with red lines that is the key to the temporal modeling of RNN.

by LSTM by introducing the forget gates (Fig. 2.2 (c)):

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
h_t &= o_t tanh(c_t)\,,
\end{aligned}
\tag{2.16}
$$

where i_t is the input gate that masks out the standard RNN inputs, f_t is the forget gate that masks out the previous cell, c_t is the cell that stores a mixture from the input and the forget gates and o_t is the output gate that masks out the values of the current hidden states h_t. The RNN, or widely-used LSTM, is a powerful structure to capture the temporal information of input features. This idea drives us to use the RNN model to enhance the weak temporal relations between the DPGMM labels [59].

Now we introduce the DPGMM-RNN hybrid model. The DPGMM-RNN hybrid model improves the DPGMM clusters in two steps. First, 1) train an RNN (with parameters W) to map (f_W) the input feature x to the given DPGMM label z. The RNN maximizes the likelihood (equivalently, minimizes the cross-entropy loss l_{ce}). We obtain an RNN-optimized parameter W^* through the gradient descent algorithm:

$$W^* = \arg\max_W \sum_i p(z_i|x_i, W)\,, \tag{2.17}$$

equivalently,

$$W^* = \arg\min_W \sum_i l_{ce}(f_W(x_i), z_i)\,. \tag{2.18}$$

2) Second, generate the DPMM-RNN label z^* by the RNN with an RNN-optimized parameter.

$$z_i^* = f_{W^*}(x_i)\,. \tag{2.19}$$

2.2.2 DPGMM-RNN Hybrid Model—Combine DPGMM and RNN

We generally construct the DPGMM-RNN hybrid model using RNN to refine the DPGMM clusters. We apply the hybrid model to the unsupervised phoneme discovery, which uses RNN to relieve the fragmentation problem of the DPGMM clusters in three steps (Fig. 2.3):

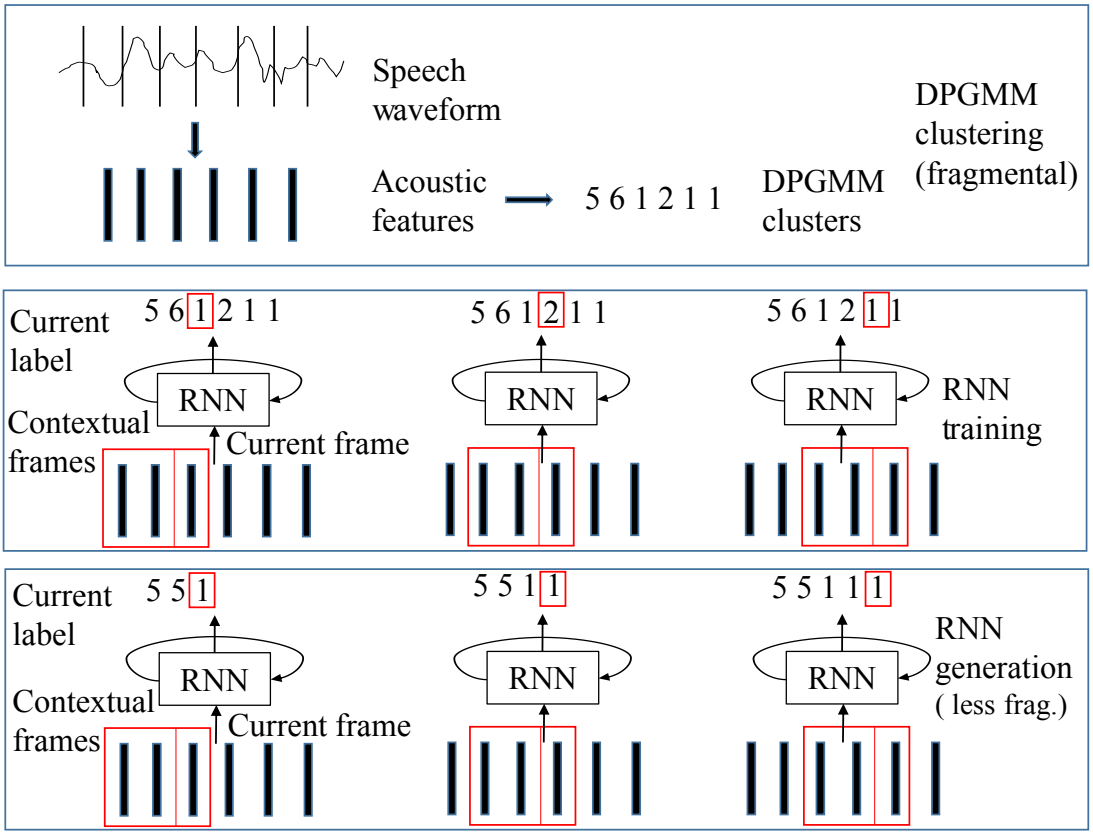


Figure 2.3: Three steps to construct DPGMM-RNN hybrid model for unsupervised phoneme discovery — DPGMM clustering, DPGMM training, and DPGMM reconstruction — using RNN to relieve segmentation problem of DPGMM clusters. The RNN target can be DPGMM cluster label for unsupervised phoneme discovery, or DPGMM posterior vector for unsupervised feature extraction.

- **DPGMM clustering**: after extracting the features from the raw audio, we apply the DPGMM clustering algorithm to get a cluster label for each feature frame. Many DPGMM segments (successive frames with identical cluster labels) are fragmental, which are much shorter (one or few frames) than phonemes in human language (Fig. 3.1). We use RNN to relieve this fragmentation problem.

- **RNN training**: we train the RNN model by mapping from a feature segment to the DPGMM cluster label (or the DPGMM posterior vector) of the last frame of that segment. RNN uses a shared memory to remember the tendency that momentarily produced the DPGMM cluster label (or the posterior vector) from the nearest acoustic segment.

- **RNN reconstruction**: we use RNN to get the posterior vector frame-wisely by inputting the speech segment. The dimension of the maximum probability in the posterior vector is chosen as the output cluster label. The RNN reconstruction of cluster labels helps relieve the fragmentation problem (Fig. 3.3). For example, DPGMM fragmental structure "a a b a a" in several successive frames is usually revised by RNN to "a a a a a."

The RNN target can be cluster labels or posterior vectors. We usually use clusters as the target if the goal is to find discrete segments for unsupervised phoneme discovery. We use posteriorgrams as the target in this paper because the goal is to find continuous features.

We explore three types of context windows: the forward window, the backward window, and the bidirectional window. For example, when using eight frames of acoustic features as RNN input context, the forward window takes a current frame along with eight past frames, the bidirectional window takes a current frame along with four past frames and four future frames, and the backward window takes a current frame and eight future frames.

2.2.3 DPGMM Based Features—Apply DPGMM Methods to ASR

We propose to extend the MFCC features with DPGMM posteriorgrams or DPGMM-RNN posteriorgrams by concatenation (Fig. 2.4) to improve ASR, where

- the MFCC features represent acoustic features,
- the DPGMM generates DPGMM posteriorgrams after adaptation of DPGMM parameters with MFCC features,

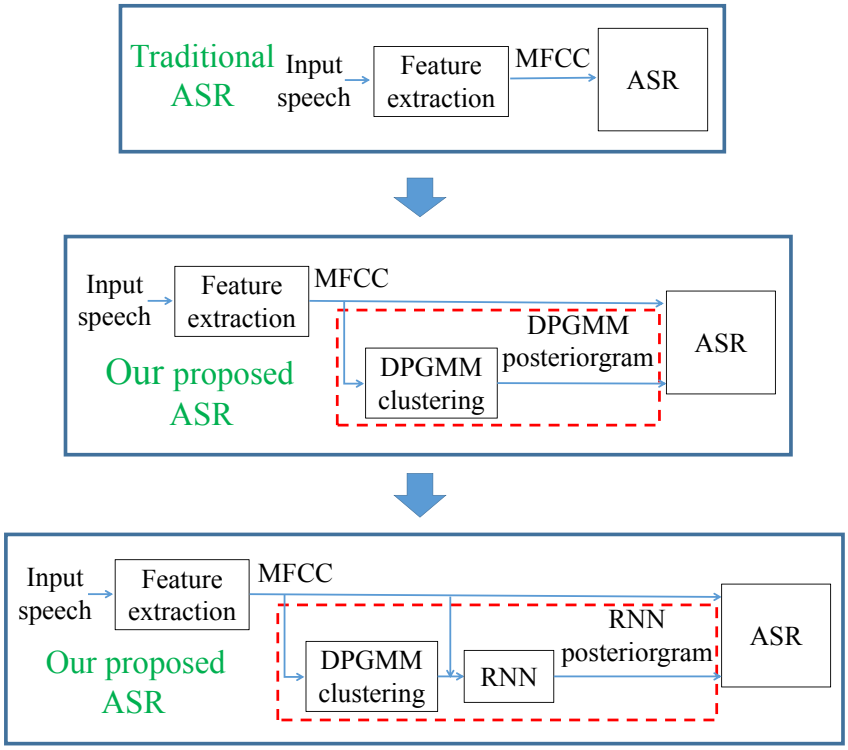


Figure 2.4: Proposed feature extension by concatenating an MFCC feature with a DPGMM posteriorgram (from the DPGMM clustering algorithm) or with an RNN posteriorgram (from the DPGMM-RNN hybrid model) in feature extraction for ASR. A posteriorgram is a vector whose k-th dimension represents the probability that an observed frame belongs to the k-th cluster.

- the DPGMM-RNN hybrid model generates DPGMM-RNN posteriorgrams (or RNN posteriorgrams for short) after adaptation of the RNN parameters to connect the MFCC chunk with the DPGMM cluster or posteriorgram,
- and the posteriorgram is a posterior probability vector whose k-th dimension represents the probability that the observed data belong to the k-th cluster.

We integrated the proposed feature extension (with a DPGMM posteriorgram or an RNN posteriorgram) into the feature extraction to improve the ASR system (Fig. 2.4).

2.3. DPGMM Based Methods to Unsupervised Phoneme Discovery and Low-resource ASR

The ASR seeks a sequence of linguistic units such as phonemes and words for each speech utterance. The state-of-art ASR system needs deep learning technology that requires a huge amount of speech and annotation resources. The ASR technology becomes mature in such rich-resource languages as European languages. Researchers attempted to develop new speech technology for the low-resource languages such as African languages with less speech recording and, more importantly, fewer linguistic annotations of the speech recordings. Such a task is meaningful but challenges the traditional deep learning technologies supervised by a huge amount of annotations.

One related task to the low-resource ASR is unsupervised phoneme discovery that attempts the find the linguistic units from an audio signal without annotation or knowledge of linguistic data. We evaluate rich-resource or low-resource ASRs by the character error rate (CER) or the word error rate (WER). The evaluation for the unsupervised phoneme discovery is more challenging. Several Zerospeech challenges [3, 29, 30] use the ABX error rates and the bitrates as evaluations. The ABX error rates [47] measure the discriminability of the representations in discriminate phonemes. Bitrates [3] in Zerospeech challenges are defined with two steps: treat each feature frame symbolically as a character. The test set of all utterances becomes a long sequence of characters, from whose frequency they

compute the character entropy. The bitrate of the test set is defined as the total entropy — the entropy-per-character multiplied by the number of characters — divided by the total duration of speech of the whole test set. Ideal representations of speech should have low ABX error rates and bitrates similar to a textual language with a low bitrate. We will examine our proposed DPGMM based unsupervised features on the tasks of unsupervised phoneme discovery and low-resource ASR.

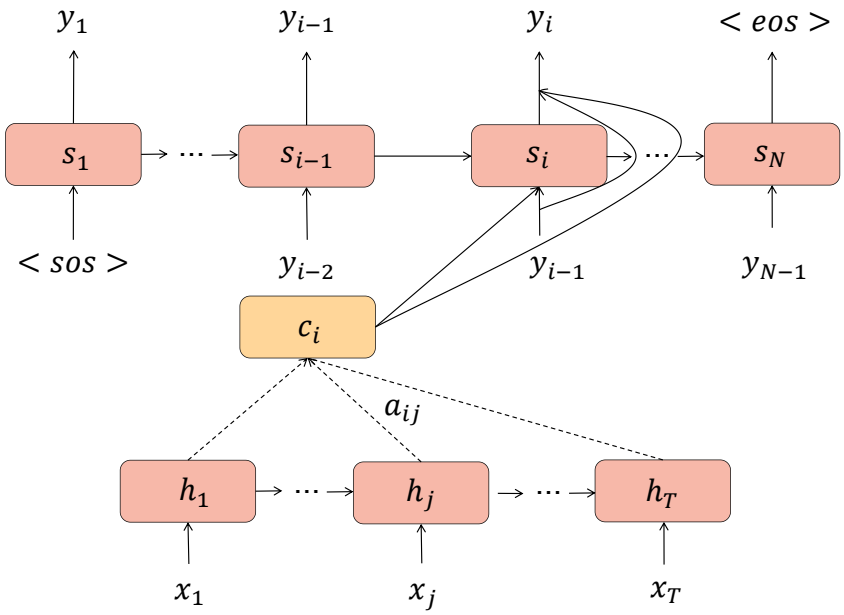


Figure 2.5: Structure of an attentional encoder-decoder ASR.

Now we describe the attentional encoder-decoder ASR model used in our experiments (Fig. 2.5). The ASR is trained with the maximum likelihood estimation. The joint probability of the likelihood is decomposed into conditional probabilities (guesses) formulated as follows. Assume that the ASR hears an utterance $x = (x_1, \ldots, x_j, \ldots, x_T)$ and attempts to guess one word (or one character) after another until it detects the termination signal (of symbol <eos>). The ASR guesses the current word y_i according to the current state s_i, the most recently guessed word y_{i-1} and the acoustic context c_i

$$p(y_i|y_1, \ldots, y_{i-1}, x) = f(s_i, y_{i-1}, c_i)\,, \tag{2.20}$$

where the current state s_i is inferred from the past state s_{i-1}, the past word y_{i-1}, and the acoustic context c_i,

$$s_i = g(s_{i-1}, y_{i-1}, c_i)\,, \tag{2.21}$$

the acoustic context c_i is summarized by ASR's attention $a = (a_{i1} \ldots, a_{ij}, \ldots, a_{iT})$ distributed on the RNN-encoded states $h = (h_1 \ldots, h_j, \ldots, h_T)$ of the utterance,

$$c_i = \sum_{j=1}^{T} a_{ij} h_j\,, \tag{2.22}$$

and the attention a_{ij} is normalized as weights whose summation is one. The attention a'_{ij} (before normalization) of the current ith guess on the jth frame depends on how similar this frame h_j to the past state s_{i-1}. Such similarity is measured by inner production similarity $a'_{ij} =< s_{i-i}, h_j >$ or neural network similarity $a'_{ij} = V^T tanh W[s_{i-i}, h_j]$, where the neural network has two layers with parameters V and W that maps the concatenated vector to a scalar similarity a'_{ij}. The whole model of encoder-decoder attentional ASR decodes (guesses) one word after another by attending to different parts of an utterance for each word.

Chapter 3

DPGMM-RNN Hybrid Model for Unsupervised Phoneme Discovery

3.1. Motivation: A Perception Driven Approach

Deep neural network technology has recently achieved great success by learning from a large amount of human annotated data. Although annotating such linguistic units as words and phonemes is essential for applying deep learning to the spoken language processing, it is expensive, time consuming, and requires expert knowledge of specific languages. One solution is to directly identify phoneme-like units from speech by machine learning (unsupervised phoneme discovery) instead of human annotation.

As we mentioned earlier in the Section 1.1.3, unsupervised phoneme discovery [38, 48] or similar tasks [49, 50, 51] have been explored by different experiment settings with different measurements. Recently the Zerospeech Challenge [29] was organized to compare the performance of these methods. Typical methods include neural network technology, such as representation learning by autoencoder [32, 33, 34] or discriminative training by ABnet [35], traditional clustering such as GMM [36] or k-means [37, 36], and nonparametric clustering such as the Dirichlet Process Gaussian Mixture Model (DPGMM) trained by Gibbs sampling [38], or variational inference [39, 40]. Among them, DPGMM, which is acoustic

clustering, achieved the top performance at Zerospeech 2015 and 2017 [52, 53].

An acoustic driven approach (e.g., DPGMM clustering) identifies different acoustic patterns and treats them as different linguistic units such as phonemes. It sometimes discovers acoustic segments that do not agree with phonemes. For example, in Japanese, /r/ and /l/ are acoustically different without distinguishing the meaning of the utterances, and thus they are treated as the same phoneme. Sometimes abrupt or local changes, such as a sudden burst of air that is released at the stop of /p/, create several acoustic segments inside one phoneme.

To tackle the problems of the acoustic driven approach, we propose an alternative perception driven approach and introduce the concept of the perception bias of phonemes (against acoustic speech) and a method to deal with it.

3.1.1 Bias of Phoneme Perception

Identifying phonemes from natural speech is challenging. Early studies on the high correlation between sound spectra and isolated phonemes provided encouragement that the problem could be solved. For example, we can identify vowels by formant values or stops by silent periods, which are verified by the speech synthesis practice [54, 55]. However, seeking phonemes from the spectrum in spontaneous speech flow is frustrating. While we are listening to some phonemes, features, or breaks at certain moments, we cannot find enough evidence about them from the spectrum [56]. The spectrum faithfully reacts to energy of different frequencies at a certain moment; it doesn't react to the sound history or subsequent sounds. However, our phoneme perception is biased. Instead of merely momentarily decoding the speech, our perception is influenced by the expectation of what will come next or our speaking and hearing experience. The lack of correspondence between speech perception and sound stream forms a central challenge in phoneme discovery from spontaneous speech [57].

The human perception of phonemes is biased against speech sounds. For example, when a virtually identical burst happens before /i/, /a/, or /u/, we tend to hear /pi/, /ka/, or /pu/ [56] because we hear them while referencing how we say them [57]. Since the lips are close together when we generate /i/ and /u/, this bilabial articulation may interpret the burst as /p/; when the tongue is relatively low and back when we generate /a/, dorso-velar articulation

may interpret the burst as /k/. Visual context can also bias our perception [58]. When hearing a recording /ba/ while watching a video of a face saying /ga/, the listeners report that they hear /da/. The compromise between visual and auditory information indicates that with accurate visual information, we can probably correct the phonemes.

Human perception has an "auto filling" ability for perceiving phonemes in sound streams [59]. Even when a phoneme (with its transition cues) is replaced by noise, people report they hear it and don't notice any noise or its location. Our lexicon knowledge influences our perception of phonemes. By adding an identical, intermediate sound between /d/ and /t/ in front of "ask" and "ash," Ganong found that people reported hearing "task" rather than "dask" and "dash" rather than "tash" [60]. Sometimes our perception relies less on lexical knowledge and more on the probability of sequences of phonemes (e.g., compensation for co-articulation varies with phonotactic probability [61]). A person's speaking and listening experiences, including the segment probabilities or sequences as well as how he says or hears these segments to achieve economical communication, also implicitly bias his perception.

Phoneme perception categorizes acoustic sounds [57], which shows another fundamental perceptual bias. If we create linearly changed acoustic stimuli between two phonemes, such as /t/ and /d/, our perception nonlinearly jumps from one category to another because we cannot identify different acoustic realizations inside one phoneme category.

The above studies show that our phoneme perception is biased. Perception bias becomes a big problem in unsupervised phoneme discovery (Zero Resource Speech Challenge [3], as we introduced at the beginning of our paper), which asks machines to learn phonemes from acoustic speech in an unsupervised way [38, 3]. A machine learning algorithm discovers *objective* acoustic segments from speech, while humans annotate *subjective* perceptual phonemes as underground truth with perception bias.

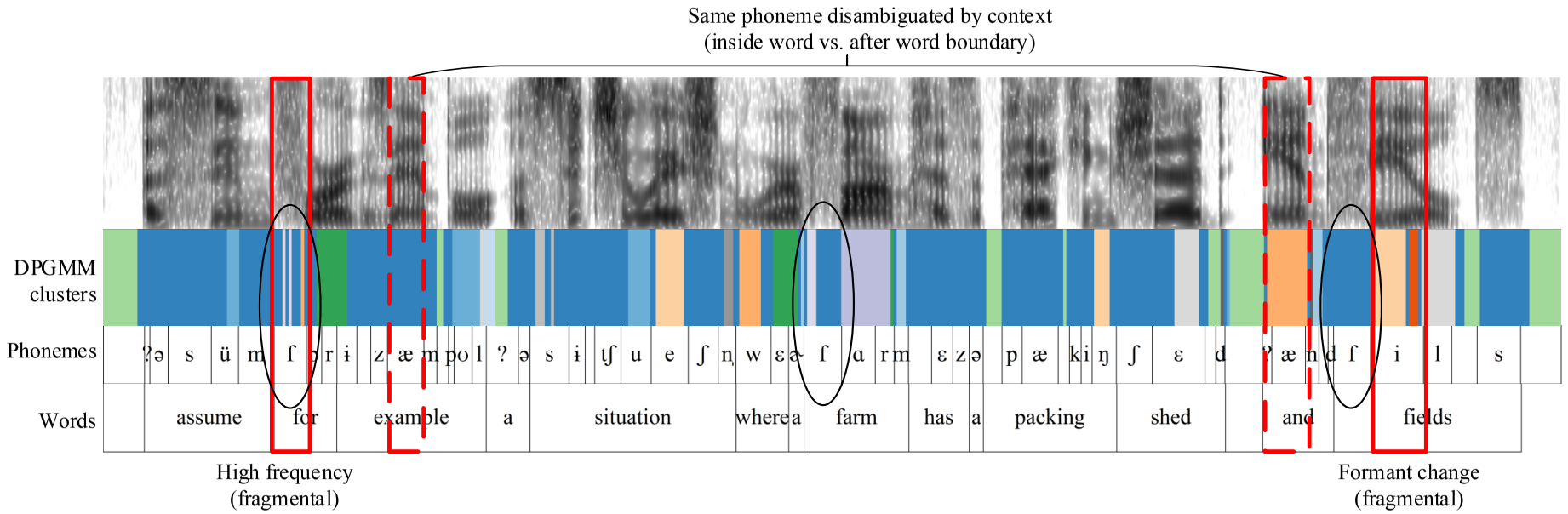


Figure 3.1: Example [1] that shows problems of DPGMM clustering in unsupervised phoneme discovery from TIMIT corpus [2]. Top layer is spectrogram followed by DPGMM label, phoneme, and word layers. In second layer, each color denotes one specific type of DPGMM cluster. Red, solid-lined rectangles show that complex acoustics such as fricatives with high frequency and vowels with rapid formant change cause fragmental DPGMM clusters; small black circles show that we can improve categorization of identical phoneme if there are no fragments; red, dash-lined rectangles show two acoustically different segments disambiguated by surroundings, which should be same phoneme, are treated as different clusters.

For example, in Fig. 3.1, the clustering algorithm treats the same phoneme /æ/ in 'example' and 'and' as different acoustic segments because their acoustical spectra are quite different; it treats the same phoneme /f/ in 'for,' 'farm,' and 'fields' as different acoustic segments by faithfully recording the acoustic fragmental realizations inside the phoneme category.

In the following sections, this paper proposes one method to tackle the disagreement between phonemes and acoustic signals caused by the perceptual bias for unsupervised phoneme discovery.

3.1.2 DPGMM-RNN Model and Phoneme Categorization

Machines can directly get discrete segments by applying such clustering algorithms as K-means [37, 36], GMM [36], or DPGMM clustering [39, 38, 40] from the acoustic features. The DPGMM algorithm [62] retained the state-of-the-art approach in the Zerospeech 2015 and 2017 [52, 53].

However, framewisely clustering acoustic features to get segments suffers from the intra-phoneme fragmentation problem (Fig. 3.1). First, these traditional clustering algorithms cannot fully capture the temporal information of speech features. As long as the spatial distribution of these acoustic features in high-dimensional space does not change, such clustering algorithms as K-means or GMM always get similar results because they ignore the time order of these features. The DPGMM algorithm introduces the Dirichlet Process (DP) to help dynamically create new clusters at every moment based on the frequency of the clusters of all the previous frames without considering their order [63]. Theoretically, DP is infinitely exchangeable; joint distribution doesn't depend on the order of data if they are infinite [64]. We believe DPGMM involves weak temporal contextual modeling for finite sequential data clustering. Second, in actual unsupervised phoneme discovery practices, after carefully tuning the parameters (e.g., DPGMM's concentration parameter, which is closely related to the number of clusters), such optimal performances (in discriminating phonemes in different languages) always create more clusters than the number of phonemes in normal human languages [1, 53]. Third, the DPGMM algorithm creates small clusters inside one phoneme—the fragmentation problem—with such a complex acoustic structure as a fricative with high-frequency components or a vowel with a

sharp format change [1]. The DPGMM algorithm tries to get higher resolution by struggling to discriminate among acoustically complex phonemes, which also tends to increase the number of clusters overall.

Human perception of speech is categorical [65]. We don't hear intra-phoneme fragmental details when discriminating complex phonemes [59, 65]. For general sounds, for example, people can discriminate about 2000 different pitches, but they can only identify about seven absolute ones [57]. For speech sounds, however, similar discriminability and identifiability of phonemes make people fail to discriminate the acoustic variations inside each phoneme category [65]. If we believe one phoneme type is a set of segments, then our biased perception cannot distinguish within the set, including the unstable fragmental acoustic realizations, created from the clustering algorithm, of these segments.

In this paper, we propose the DPGMM-RNN hybrid model, which uses RNN to relieve the DPGMM intra-phoneme fragmentation problem. Assume that human perception is hierarchical at low-level perception. At the first run, very low-level, bottom-up unbiased clustering can get fragmental details with sufficient discriminability of the segments from the raw stimuli of speech by air vibration. At the second run, the ear uses a primitive top-down acoustical contextual refinement and pays little attention to the variations inside one phoneme. Such perceptual refinement can be achieved by RNN mapping (Fig. 2.3). We train RNN intensively by remembering the phoneme (DPGMM clusters) at every moment of speech by listening to a chunk of sound that includes that moment. Listening by chunks helps integrate long acoustical contexts as a whole instead of concentrating on random short-time fragmental changes over a few frames. After RNN remembers different chunk realizations for each phoneme, it has the ability to identify the sounds. We show that RNN refinement relieves the fragmentation problem inside phoneme categories.

Moreover, facing the weak contextual modeling of DPGMM, whose joint likelihood does not depend on the order of the observed data when they are infinite [64] and mainly captures acoustic information at the frame level, RNN refinement with chunks of successive frames instead of single frames of speech can rediscover such temporal structures as formant transitions, which cross several frames and are important acoustic cues to perceptually discriminate phonemes in spontaneous

speech [57].

In human perception, top-down contextual constraints, not only acoustical or phonemic ones but also linguistic and lexical ones help correct or remove the segments that are wrongly discovered to make them closer to phoneme units [60]. In previous research [66, 67, 68] on infant phoneme acquisition, the words and phonemes are assumed to be acquired at the same period and jointly optimize each other, supported by phoneme-lexicon joint discovery by adaptor grammar framework [69] and hierarchical nonparametric Bayesian model [70]. Language model of discovered segments was trained to optimize phoneme discovery [71]. In our proposed DPGMM-RNN hybrid model, RNN remembers the statistical structure, reflecting on such contextual constraints at the phonetic and lexicon levels, of audio segments from DPGMM clusters. For example, if the DPGMM clustering algorithm confuses 'bite' with 'kite' inside an utterance that contains the sound 'dog,' RNN can correct such mistakes because sounds 'dog' and 'bite' are semantically correlated. RNN remembers their co-occurrence.

We propose the DPGMM-RNN hybrid model for decoding segments from speech signals. The DPGMM algorithm finds fragmental segments, while RNN fixes the fragmentation problem. From the view of using machine learning to track the bias of human perception, our DPGMM-RNN hybrid model achieves better phoneme categorization by solving the fragmentation problem.

We propose to use conditional perplexity (of clusters conditioned on phoneme) to measure the fragmental level of the discovered segments. We use the conidtional entropy to evaluate our proposals, which is the average number of clusters per phoneme; we also use the ABX discriminability score [31] to evaluate our proposals, which is the cluster represenation's ability in discriminating among phonemes.

3.2. Experiments

3.2.1 Evaluation Metric

We evaluated our generated segments with conditional entropy-based measurements (conditional perplexity, homogeneity, completeness, and v-measure) and

psychology-based measurements (ABX discriminability score and ABX error rate).

We proposed the DPGMM-RNN hybrid model to relieve the fragmentation problem of DPGMM clusters. To measure the fragmental level of the generated representation, we computed the average number of cluster types corresponding to one phoneme type, **conditional perplexity**, with the exponential of the conditional entropy of cluster C conditioned on phoneme truth T with base 2:

$$ppl(C|T) = 2^{H(C|T)}\,, \tag{3.1}$$

$$\begin{aligned} H(C|T) &= \sum_t p(t) H(C|T=t) \\ &= -\sum_t p(t) \sum_c p(c|t) \cdot log p(c|t) \\ &= -\sum_t \frac{n_t}{n} \sum_c \frac{n_{ct}}{n_t} \cdot log \frac{n_{ct}}{n_t}\,, \end{aligned} \tag{3.2}$$

where n is the number of frames, n_t is the number of frames of phoneme truth t, and n_{ct} is the number of frames annotated as phoneme t and clustered as cluster c.

However, the conditional perplexity is insufficient to describe the matching degree of the generated clusters and the underground phonemes. For example, if we generate identical clusters for the whole corpus, which means that no fragments exist. Then conditional perplexity of cluster given phoneme is the lowest, however, the discovered identical clusters mismatch the different phonemes.

In another word, the conditional perplexity detects an oversegmention problem that one phoneme has several cluster segments inside, but it ignores the undersegmention problem that one cluster segment covers several phonemes. Besides the amount of clusters per phoneme class (the conditional perplexity), we also need the amount of phoneme classes per cluster as an additional measurement.

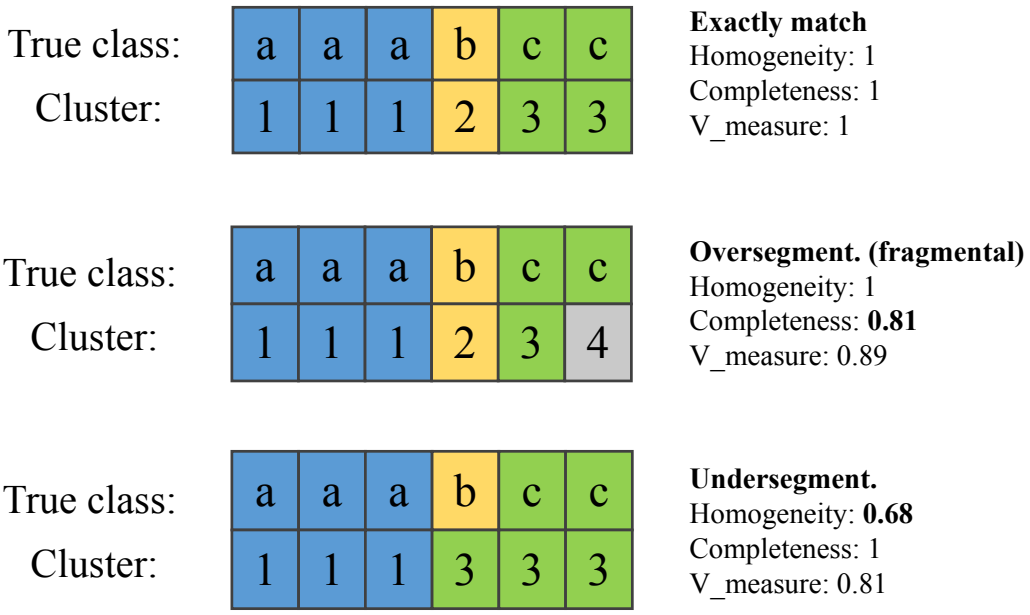


Figure 3.2: Relationship between evaluation metrics (homogeneity, completeness, and v-measure) and mismatching problems (fragmentation, oversegmentation, and undersegmentation). All metrics range between 0 to 1; higher value means better matching.

Completeness, **homogeneity**, and **v-measure** (Fig. 3.2) are measurements [72] (similar to accuracy, recall, and F-score) that reflect the matching degree between generated clusters and underground phonemes using normalized conditional entropy. Completeness c, homogeneity h, and v-measure v (harmonic mean of h and c) are defined as follows:

$$c = 1 - \frac{H(C|T)}{H(C)}\,, \tag{3.3}$$

$$h = 1 - \frac{H(T|C)}{H(T)}\,, \tag{3.4}$$

$$v = \frac{c \cdot h}{c + h}\,. \tag{3.5}$$

We compute the entropy and the conditional entropy by the relative frequency, similar to Eq. (3.2) of the framewise samples from generated cluster C and

phoneme truth T. All three measurements were normalized between 0 to 1 (as $H(T) \geq H(T|C)$), and the higher value shows better matching between the generated clusters and the underground phonemes. High completeness shows that each phoneme type almost 'completely' (completeness) corresponds to a unique generated cluster type; high homogeneity shows that one cluster type should correspond to the 'same' (homogeneity) phoneme truth type.

As shown in Fig. 3.2, low completeness indicates that the cluster representation is fragmental or oversegmental with respect to the phoneme truth. Low homogeneity indicates undersegmental. Only high v-measure indicates that the representation is neither oversegmental (fragmental) nor undersegmental because the clusters agree with the phonemes.

In addition to the above conditional entropy-based measurements, which are based on the global frequency, we also evaluate our representation by the discriminability of the local phoneme segments using psychological measurements: ABX discriminability score (or its reverse: ABX error rate) [31].

In auditory perception experiments, we used the ABX test to measure a subject's ability to discriminate between sound categories A and B. The subject hears sound A, then sound B, and finally a third sound X that is either from category A or category B. Here we assume X belongs to category A. If the perception distance between A and X is less than that of B and X, then the subject will think sounds X and A are from the same category, which indicates he can discriminate between category A and category B.

If we replace the subjective perception distance with the objective distance of our cluster representation (e.g., the cosine distance between the one-hot representation of the clusters), then the ABX test can measure the ability of the clusters to discriminate among different sound segments: **ABX discriminability score**. For example, given triphone a-p-a as A, another triphone a-b-a as B, and a third triphone a-p-a as X, based on a cluster presentation of triphones, if the distance between A and X is smaller than that between B and X, then the ABX discriminability score of the triplet (A, B, X) is +1, and otherwise the ABX discriminability score is -1.

In theory, we can also define the discriminability score between the triphone category pair ($c1$, $c2$) [47] by taking samples A and X from $c1$ and sample B from

$c2$ and define score s and its point estimator as follows:

$$\begin{aligned} s(c1,c2) &= p(d(A,X) < d(B,X)|A, X \in c1, B \in c2) \\ &+ \frac{1}{2}p(d(A,X) = d(B,X)|A, X \in c1, B \in c2) \\ &= \frac{1}{m(m-1)n} \sum_{a \in c1} \sum_{x \in (c1-\{a\})} \sum_{b \in c2} \\ &(\delta_{d(a,x)<d(b,x)} + \frac{1}{2}\delta_{d(a,x)=d(b,x)}) \,, \end{aligned} \tag{3.6}$$

where δ_c is the indicator function (taking value 1 if condition c is true and 0 if c is false). Coefficient m is the number of triphones of the $c1$ category, and n is the number of triphones of the $c2$ category. Metric d is any distance of the cluster representation between triphone segments, which are extracted by the phoneme annotation. We computed three specific distances between the triphone segments, with possible different number of frames, by Dynamic Time Warping (DTW) based on a frame-to-frame cosine distance (cos), symmetric Kullback-Leibler divergence (kl), and edit distance (edit) [47].

We computed the frame-to-frame distances between two frame feature vectors $x = (x_1, \ldots, x_D)$ and $y = (y_1, \ldots, y_D)$ with identical dimension D according to the following equations:

$$d_{cos}(x,y) = \frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{D} x_i y_i}{\sqrt{\sum_{i=1}^{D} x_i^2}\sqrt{\sum_{i=1}^{D} y_i^2}} \,, \tag{3.7}$$

where $d_{cos}(x,y)$ is the cosine distance between the two features and $|x|$ and $|y|$ are their magnitudes.

$$\begin{aligned} d_{kl}(x,y) &= \frac{1}{2}KL(x||y) + \frac{1}{2}KL(y||x) \\ &= \frac{1}{2}\sum_{i=1}^{D} x_i log\frac{x_i}{y_i} + \frac{1}{2}\sum_{i=1}^{D} y_i log\frac{y_i}{x_i} \,, \end{aligned} \tag{3.8}$$

where $d_{kl}(x,y)$ is the symmetric Kullback-Leibler divergence between the two features and $KL(x||y)$ is the Kullback-Leibler divergence. Note that here the feature $x = (x_1, \ldots, x_D)$ should be a distribution under the constraint that $\sum_{i=1}^{D} x_i = 1$; the feature y also should follow the constraint.

$$d_{edit}(x,y) = d_{D,D}(x,y)$$

$$d_{i,j}(x,y) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} d_{i-1,j}(x,y) + 1 \\ d_{i,j-1}(x,y) + 1 \\ d_{i-1,j-1}(x,y) + \delta_{x_i \neq y_j} \end{cases} & \text{otherwise,} \end{cases} \tag{3.9}$$

where $d_{edit}(x,y)$ is the edit distance between strings $x_1x_2...x_D$ and $y_1y_2...y_D$, notation $d_{i,j}(x,y)$ is the edit distance between $x_1x_2...x_i$ and $y_1y_2...y_j$, and δ_c is an indicator function (taking value 1 if condition c is true and 0 if c is false). Note that we assume the features take a binary value at each dimension.

The **ABX error rate** [47] is defined as one minus the average of the discriminability scores of all the category pairs with corresponding triplets A, B and X. If (A, B, X) comes from the same speaker, we call it the ABX error rate within speakers. If (A, X) and (B, X) come from different speakers, then we call it the ABX error rate across speakers.

3.2.2 Dataset and Experiment Setup

Dataset

We analyzed the DPGMM-RNN hybrid model with the test set of the TIMIT corpus [2], which contains 0.81 hours read speech with 1344 English utterances.

We compared the DPGMM-RNN hybrid model with the methods that achieved the top results in Zerospeech 2019 [3] with English read speech: 5941 training utterances spoken by 100 speakers (about 15 hours and 40 minutes) and 455 test utterances spoken by 24 speakers (about 28 minutes).

Experiment setup

We used 39-dimensional MFCC+Δ+$\Delta\Delta$ acoustic features (25-ms frame size and 10-ms frame shift) with mean and variance normalization and vocal tract length normalization.

We obtained clusters with the DPGMM algorithm using the same parameter setting as previous works [53, 73] with a toolkit [62]. We set the concentration

parameter to 1 and the mean and variance of the prior to the global mean and the global variance of the MFCC features with belief-strengths 1 and 42. We got cluster labels after 1500 sampling iterations.

Our DPGMM-RNN hybrid model uses clusters from the DPGMM algorithm as targets. We used an RNN that contains 3 layers of LSTM with input layer and hidden layer sizes of 39 and 512, and output layer size matching the number of DPGMM clusters. The training of RNN uses 20 epochs with a batch size of 256.

We trained RNN from discrete DPGMM cluster labels with cross entropy loss, denoted as "RNNLabel;" we also trained RNN from continuous DPGMM posteriorgrams with MSE loss, denoted as "RNNPost." We experimented with RNN with different contexts. First, we explored the length of the context with "RNNn" that denotes the DPGMM-RNN hybrid model with an RNN trained with n past contextual frames with cross entropy loss. Second, we explored the directions of the context: "RNN_forward," "RNN_backward," and "RNN_bidirectional." For example, when using the eight frames of acoustic features as RNN input context, "RNN_forward" takes a current frame along with the eight past frames, "RNN_bidirectional" takes a current frame along with four past frames and four future frames, and "RNN_backward" takes a current frame and the eight future frames.

We got the conditional entropy-based measurements (conditional perplexity, completeness, homogeneity, and v-measure) by python and Scikit-learn [74]. We computed the ABX discriminability scores and the ABX error rates with a toolkit provided by Zerospeech 2015 and compared the DPGMM methods with other methods proposed in Zerospeech 2019 with its official evaluation program.

3.3. Results

3.3.1 DPGMM-RNN Hybrid Model and Fragmentation Problem

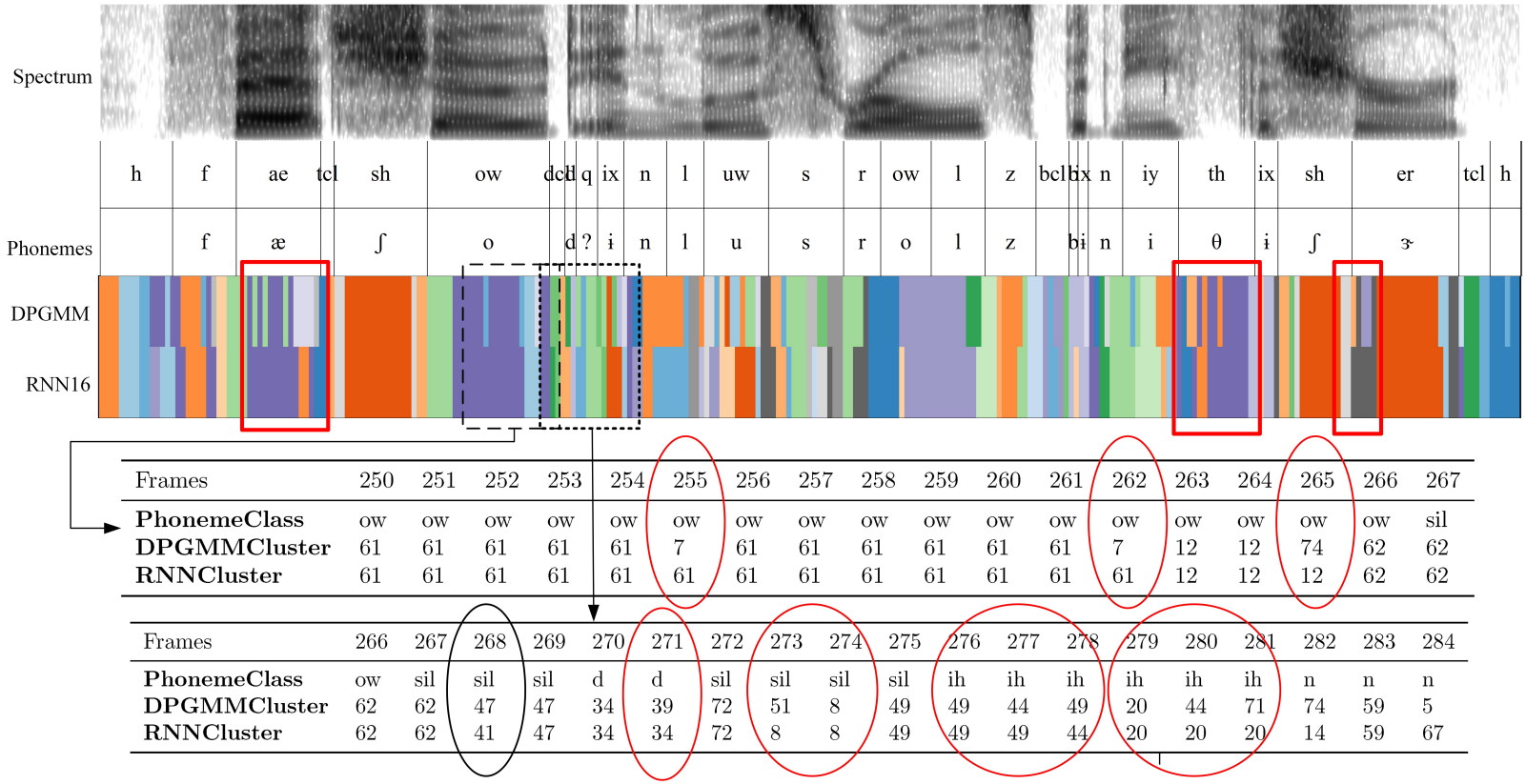
We first use two examples to illustrate how the DPGMM-RNN hybrid model relieved the fragmentation problem and later demonstrate the quantitative metrics of the fragmentation level, such as conditional entropy and completeness, in Sections 3.3.2 and 3.3.3.

Figure 3.3 shows that the DPGMM algorithm generates fragmental segments inside phonemes, and some fragments disappear after applying RNN reconstruction (as shown by red circles).

Figure 3.4 shows that the DPGMM-RNN hybrid model decreases tiny fragments using RNN by accepting longer chunks, each of which is composed of the current frame and its past successive frames, with stronger contextual modeling.

Table 3.1: Mapping from distinctive feature to phonemes. We represent phonemes as 39 TIMIT phonemic symbols used by the Kaldi recipe [5]. We also included International Phonetic Alphabet (IPA) of TIMIT phoneme symbols. 'stop_u' denotes an unvoiced stop; 'stop_v' denotes a voiced stop; 'ʃ(ʒ)' means ʃ and ʒ are represented as identical TIMIT phonemic symbol 'sh'.

Feature	Phoneme	IPA
Stop_u	p t k	p t k
Stop_v	b d g	b d ɡ
Affricate_u	ch	tʃ
Affricate_v	jh	dʒ
Fricative_u	hh f th s	h f θ s
Fricative_v	sh v dh z	ʃ(ʒ) v ð z
Nasal	m n ng	m n ŋ
Semivowel	w l r y	w l r y
Diphthong	ay oy aw ey ow	aɪ ɔɪ au eɪ əu
Front	iy ih eh ae	iː ɪ e æ
Mid	er ah aa	ɜː ʌ(ə) ɑː
Back	uw uh	uː u
Closure	dx sil	(closure) (silence)



Frames	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267
PhonemeClass	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	ow	sil
DPGMMCluster	61	61	61	61	61	7	61	61	61	61	61	61	7	12	12	74	62	62
RNNCluster	61	61	61	61	61	61	61	61	61	61	61	61	61	12	12	12	62	62

Frames	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284
PhonemeClass	ow	sil	sil	sil	d	d	sil	sil	sil	sil	ih	ih	ih	ih	ih	ih	n	n	n
DPGMMCluster	62	62	47	47	34	39	72	51	8	49	49	44	49	20	44	71	74	59	5
RNNCluster	62	62	41	47	34	34	72	8	8	49	49	49	44	20	20	20	14	59	67

Figure 3.3: Utterance 'Fat showed in loose rolls beneath the shirt' with id FADG0_SI1909 from TIMIT test set to show neural network helps solve fragmentation problems. Top layer is spectrogram followed by phoneme layer, DPGMM, and RNN layers. RNN16 is short for DPGMM-RNN hybrid model with 16 contextual frames. Red rectangles show neural network reduces fragmental problem; table shows more details of agreement of classes and clusters in two segments (red circles for better cases in relieving fragmentation problem, black ones for worse cases).

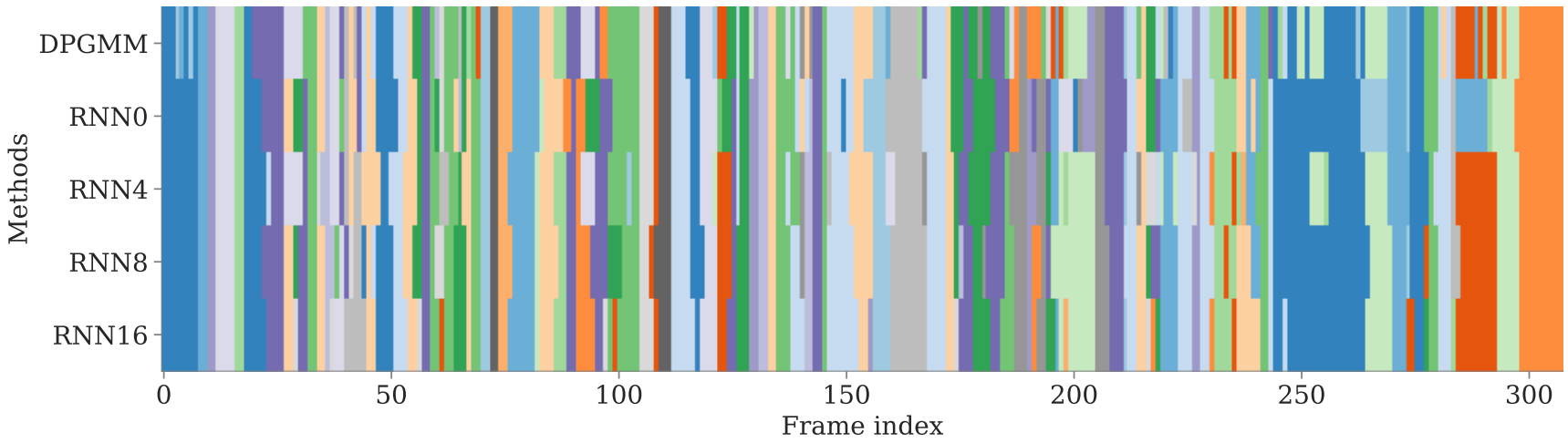
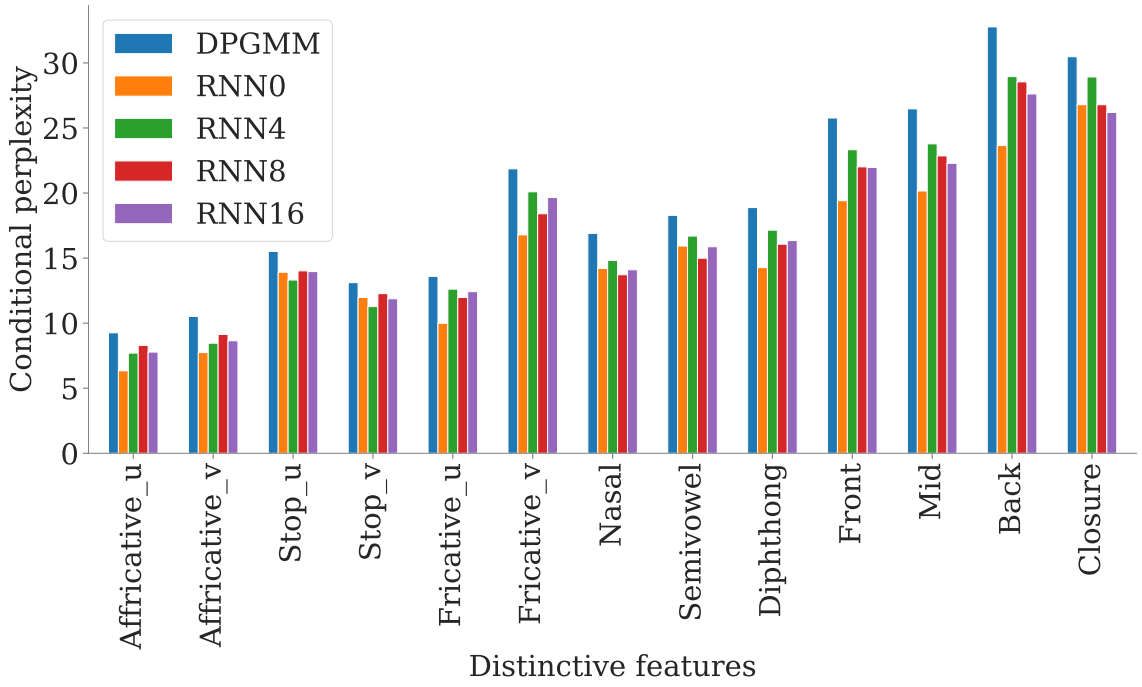


Figure 3.4: Utterance example from TIMIT test set to show fragmental level of segments decreases by applying stronger contextual modeling of DPGMM-RNN hybrid model; RNNn denotes DPGMM-RNN hybrid model with n contextual frames as RNN input.

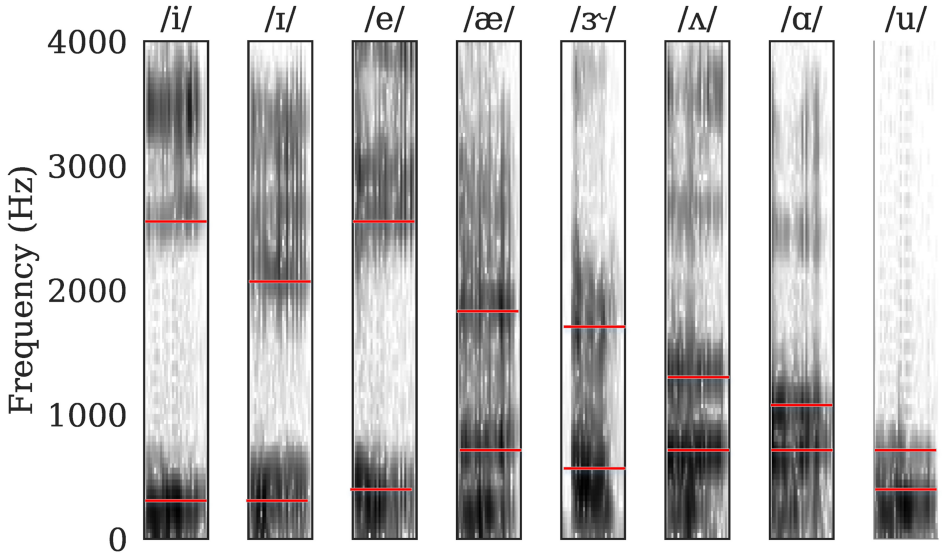
3.3.2 Distinctive Features and Fragmentation Problem

We explored why the fragmentation problem happens, how the sensitivity of the acoustics of the DPGMM clustering algorithm is associated with distinctive features, and how the proposed DPGMM-RNN hybrid model reacts to different distinctive features. After partitioning the set of phonemes into groups by distinctive features, we computed the conditional perplexity for each phoneme group to determine the average number of clusters per phoneme (the fragmental level) for each distinctive feature. Fig. 3.5a shows the following results.

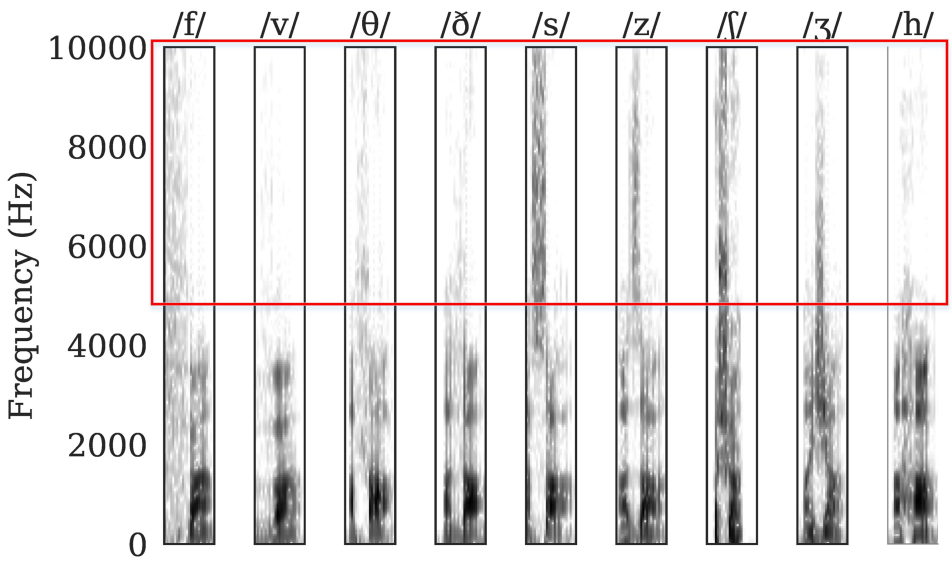
- The vowels are more fragmental than the consonants; voiced consonants are more fragmental than unvoiced ones.
- The vowels from front to back became more and more fragmental when the first and second formants became closer and harder to differentiate (Fig. 3.5b).
- The fricatives are the most fragmental consonants, which involve high-frequency components in the speech signals (Fig. 3.5c) and irregularity and rapid changes of acoustics.
- The DPGMM-RNN hybrid model (RNNn) can relieve the fragmentation problem of the DPGMM clusters (DPGMM) by decreasing the conditional perplexity for each distinctive feature.
- We computed the relative decrease ratio of the conditional perplexity between DPGMM and RNN16. Features that are more fragmental decreased more, except for affricatives and nasals.
- Even after applying RNN to relieve the fragmental problem, the conditional perplexity, which is the average number of clusters per phoneme, remained high for each feature.



(a) Conditional perplexity of clusters given phoneme classes



(b) Spectrogram of vowels from front to back with weaker discrimination of formants



(c) Spectrogram of fricatives with high-frequency noisy components

Figure 3.5: Upper subfigure (a): conditional perplexity to show fragmental level for each distinctive feature Table 3.1); RNNn denotes DPGMM-RNN hybrid model with n contextual frames. Middle subfigure (b): spectrogram of vowels from front to back; first and second formants are marked by red bars. Lower subfigure (c): spectrogram of fricatives. We extended highest frequency from 4000 to 10000 Hz compared to subfigure (a) to see high-frequency components of fricatives (inside red rectangle).

The DPGMM-RNN hybrid model can relieve the fragmentation problem and find more acoustic stable phonemes. But the fragmentation problem is far from being solved. According to the analysis of the conditional entropy, most phonemes still have more than five DPGMM-RNN clusters corresponding to them. Many DPGMM-RNN segments are short and can not cover a whole phoneme.

Overall Performance

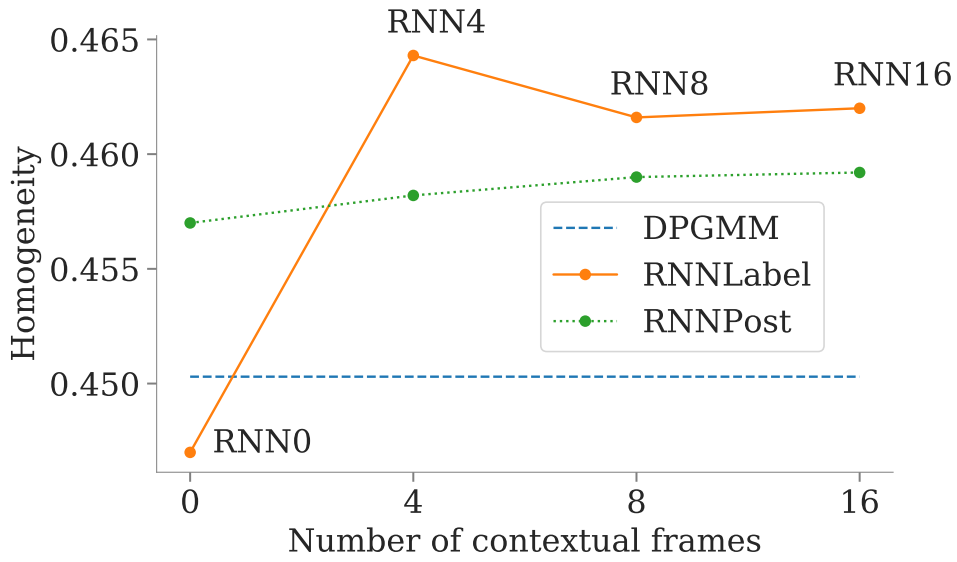
Figure 3.6 shows that the DPGMM-RNN hybrid models (RNNLabel, RNNPost) outperformed the DPGMM algorithm (DPGMM) for homogeneity, completeness, and v_measure.

Since direct RNN learning from the discrete DPGMM label always gets better results than from the continuous DPGMM posteriorgram, in later experiments, our hybrid model learned from the label (RNNn or RNNLabel) by default.

Contextual Modeling

As we increase the length of the context of the DPGMM-RNN hybrid model, the v_measure becomes larger (Fig. 3.6), showing better matching of the generated model clusters and the underground phoneme classes.

The DPGMM-RNN hybrid model utilizes the RNN to rediscover the hidden statistical structure of the speech under the supervision of the noisy DPGMM clusters. The hybrid model can correct the DPGMM cluster labels even with 0 contextual frame (RNN0), because the RNN always classifies each acoustic frame by choosing the most likely DPGMM cluster label with the maximum probability, such that the RNN correctly relabels some fragmental DPGMM clusters with extremely low probabilities from Gibbs sampling given the acoustic features.



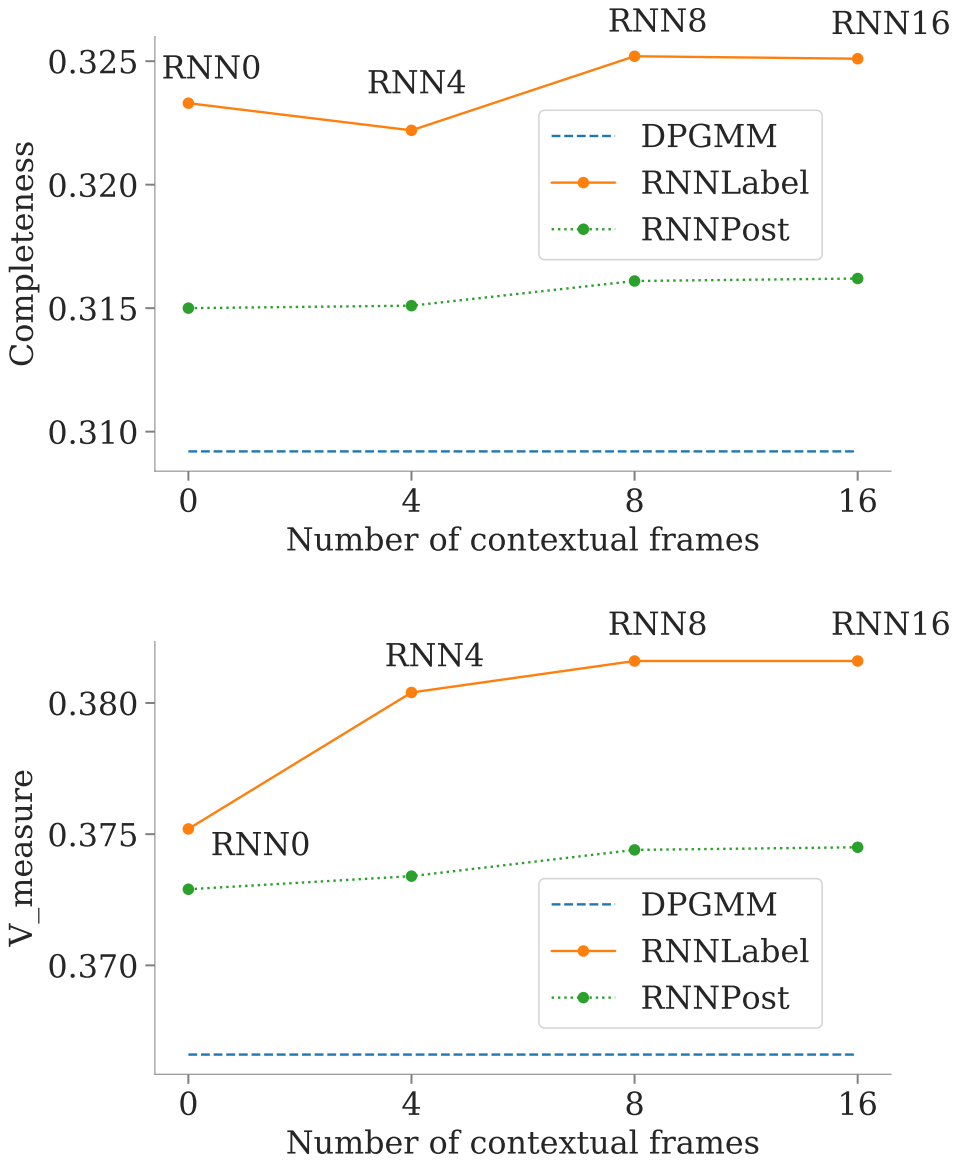


Figure 3.6: Homogeneity, completeness, and v_measure scores of TIMIT test set to show matching degree between clusters and phonemes. Dashed line is DPGMM clustering scores, and solid and dotted lines are DPGMM-RNN hybrid model scores. RNNLabel learns from discrete DPGMM cluster label with cross entropy loss; RNNPost learns from the continuous DPGMM posteriorgram with MSE loss. RNNn denotes DPGMM-RNN hybrid model with n contextual frames.

3.3.3 Analysis of Cluster Agreement with Phoneme Class

Figure 3.8 shows that the hybrid model gains better v_measure by learning the RNN from both past and future acoustic features (RNN_bidirectional) compared to merely learning from the past (RNN_forward or RNNn) or the future (RNN_backward). Since the implementation of simpler models needs less effort and makes it easier for communities to reproduce our results, most DPGMM-RNN hybrid models of this paper used the simplest strategy: learning mapping from past acoustic features (RNNn).

Oversegmentation and Undersegmentation

RNN0 (the hybrid model without a contextual frame for RNN) has relatively low homogeneity and high completeness (Fig. 3.6), which suffers from the possible undersegmentation problem: the number of cluster types of RNN0 is lower than the others (Fig. 3.7). RNN4 (hybrid model with four contextual frames for RNN) has relatively high homogeneity and low completeness (Fig. 3.6), which suffers from the possible oversegmentation problem: the number of RNN4 cluster types is higher than the other hybrid models (Fig. 3.7). Both homogeneity and completeness increase from RNN8 to RNN16.

Representation Compression

Figure 3.7 shows that the DPGMM-RNN hybrid model (RNNn) generates fewer cluster types than the DPGMM algorithm (DPGMM) and can compress the DPGMM clusters by ignoring the unstable ones with low probabilities, which makes the number of generated clusters nearer to the numbers of phonemes of the normal human languages.

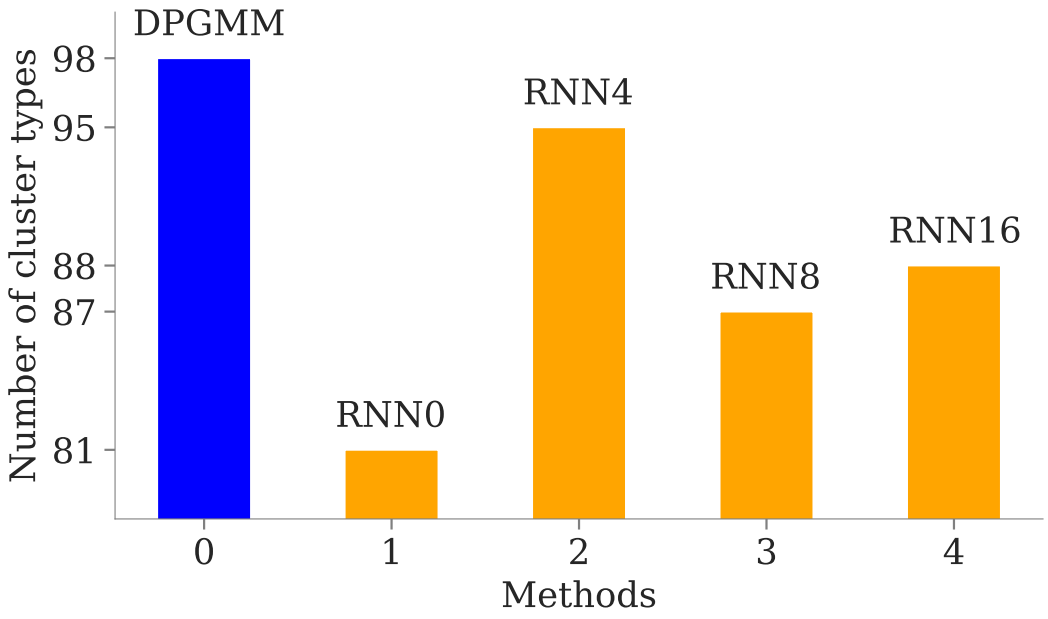


Figure 3.7: Number of cluster types from DPGMM clustering (blue bar) and that from DPGMM-RNN hybrid models with 0, 4, 8, and 16 contextual frames (orange bars).

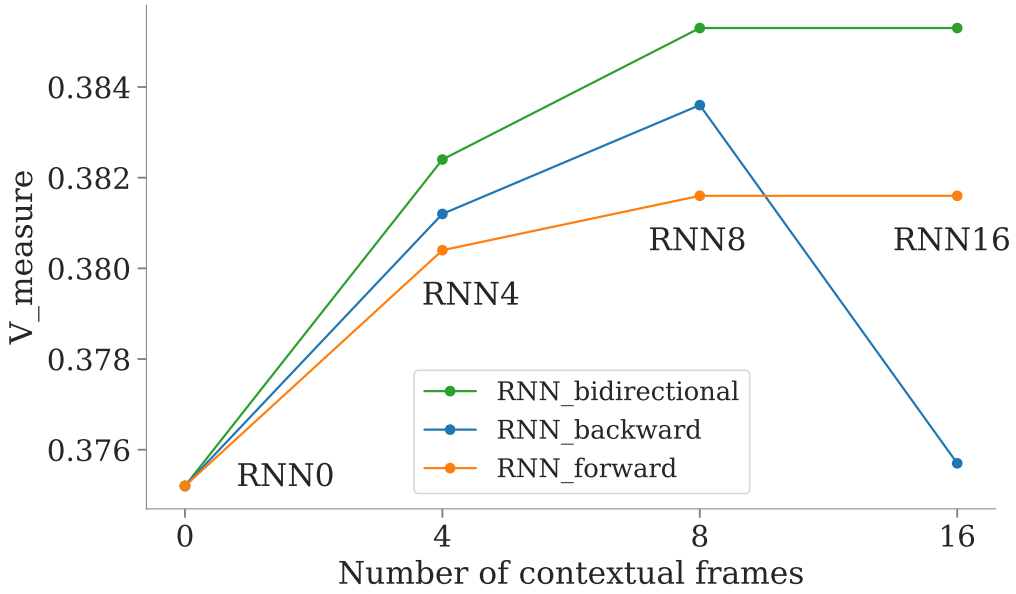
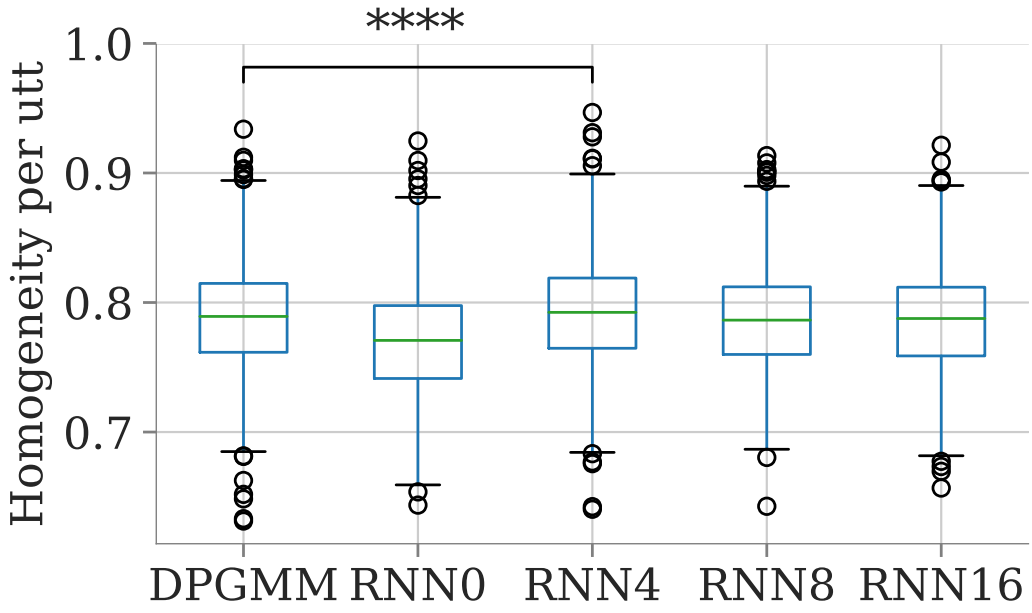
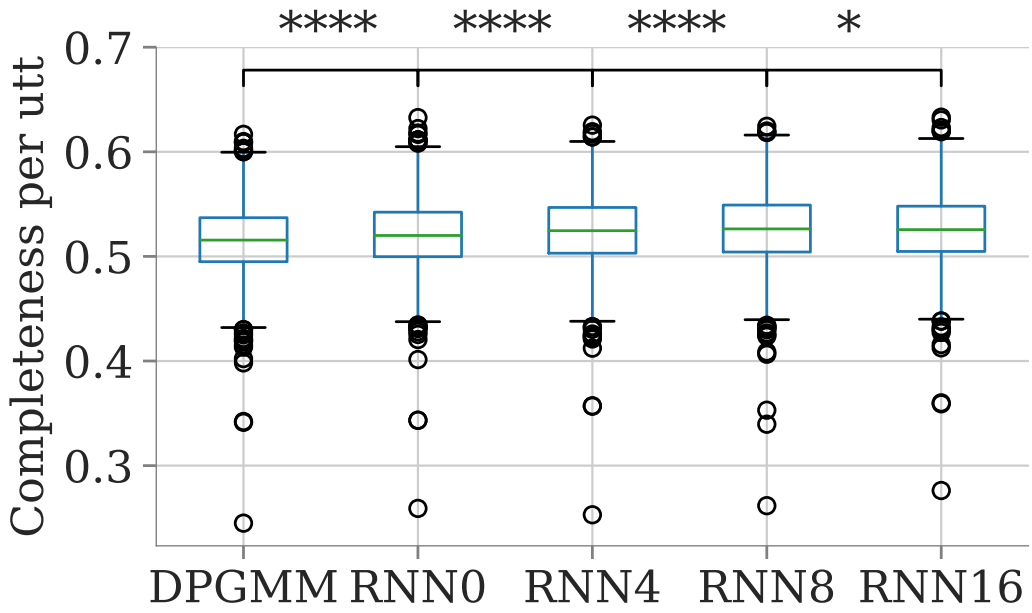


Figure 3.8: v_measure of DPGMM-RNN hybrid model with different context models. For example, when using eight frames of acoustic features as RNN input context, RNN_forward takes a current frame along with eight past frames, RNN_bidirectional takes a current frame along with four past frames and four future frames, and RNN_backward takes a current frame and eight future frames.

Performance per Utterance

Besides the above comparisons between the v_measures of the whole corpus, we also did paired t-tests on the v_measures of the utterances of the timit test set. Except for the DPGMM-RNN hybrid model with 0 contextual frames (RNN0), all the other hybrid models with longer contexts (RNN4, RNN8, and RNN16) significantly outperformed the DPGMM algorithm (DPGMM) on v_measures with p-value $p \leq 0.0001$.


****
1.0
0.9
0.8
0.7
Homogeneity per utt
DPGMM
RNN0
RNN4
RNN8
RNN16


****
****
****
*
0.7
0.6
0.5
0.4
0.3
Completeness per utt
DPGMM
RNN0
RNN4
RNN8
RNN16

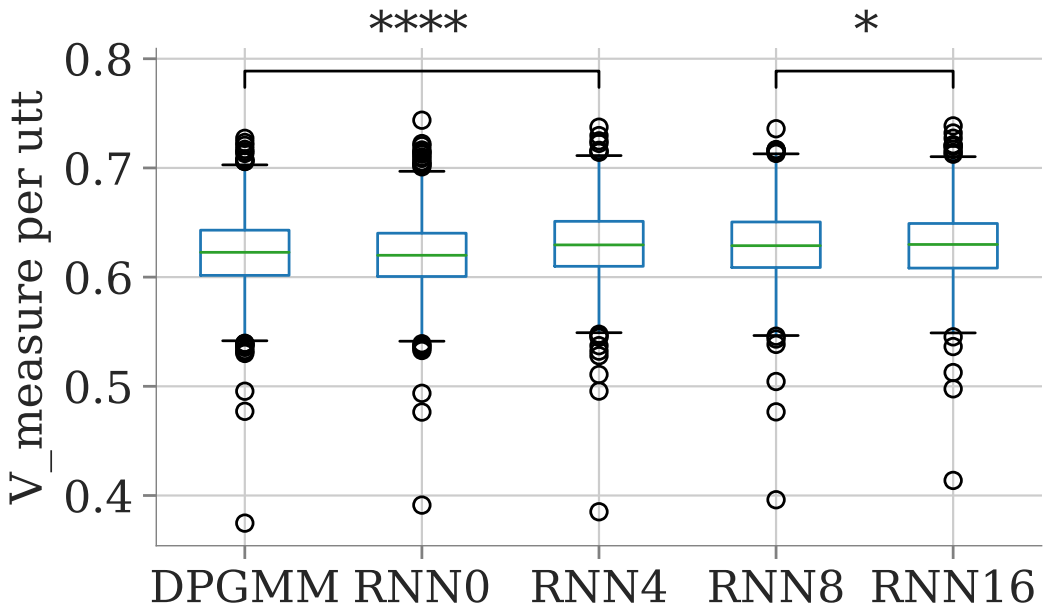


Figure 3.9: Boxplots of homogeneity, completeness and v_measure of utterances of the timit test set using DPGMM clustering algorithm and DPGMM-RNN hybrid model. RNNn is short for the DPGMM-RNN hybrid model with n contextual frames. We use the paired t-test for measures of all utterances to get the p values with the significant star **** meaning $p \leq 0.0001$ and * meaning $p \leq 0.05$.

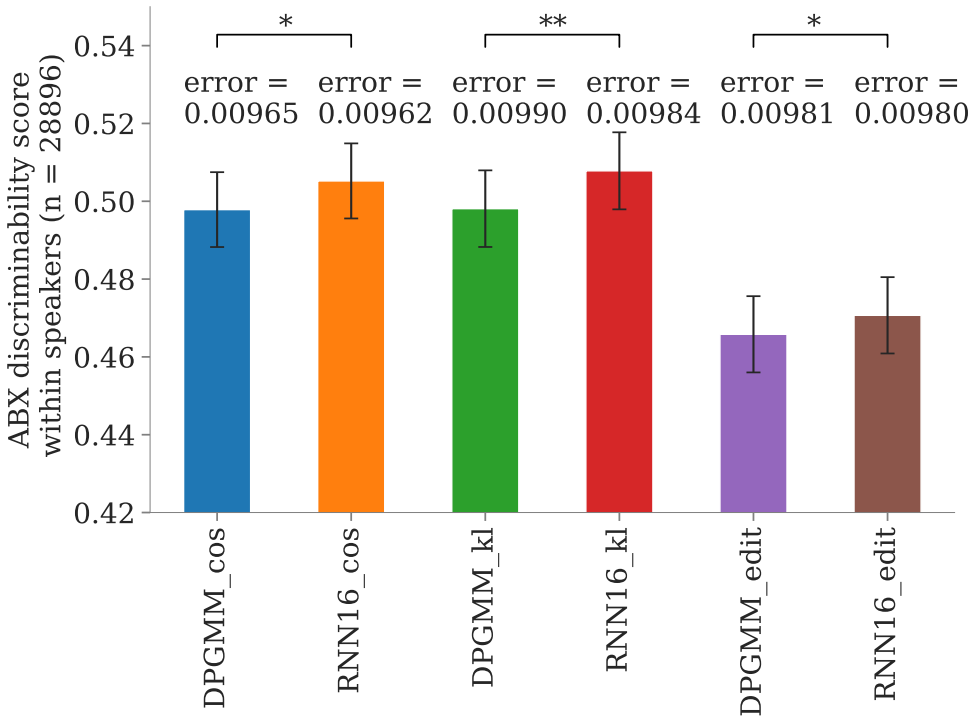
To show the statistically significant improvement of our methods, we do the paired t-test on homogeneity, completeness and v_measure of utterances of the timit test set, along with the boxplots representing their distributions (Figure 3.9). The DPGMM-RNN hybrid model optimizes to solve the fragmentation problem, because RNN maps in the direction from the acoustic signal, with hidden information of true phonemes, to the DPGMM clusters, such that intuitively each true phoneme segment should become more certain about its cluster representation. So the completeness, representing the inverse of the fragmentation level, significantly increases from DPGMM to DPGMM-RNN hybrid model, and from the DPGMM-RNN hybrid model with 0 contextual frame (RNN0) to that with 16 contextual frames (RNN16). For homogeneity and v_measure, we

find significant improvement from DPGMM clustering to DPGMM-RNN hybrid model, but slight improvement in increasing the number of contextual frames.

3.3.4 Analysis of Cluster Discriminability of Phoneme Categories

As well as the information theory inspired by measures based on the relative frequency at the global corpus level, we measured the ability of our generated clusters for discriminating the triphone categories by computing the ABX discriminability scores [29] at the local segmental level.

Figure 3.10 shows that the clusters from our proposed DPGMM-RNN hybrid model more effectively discriminate the phonemes than those from the DPGMM algorithm in ABX discriminability scores with three distances across and within speakers. The performance improvement shows statistical significance with the paired t-test. The error bar of the 95% confidence interval shows that the hybrid model achieved fewer errors than the DPGMM algorithm in ABX discriminability scores.



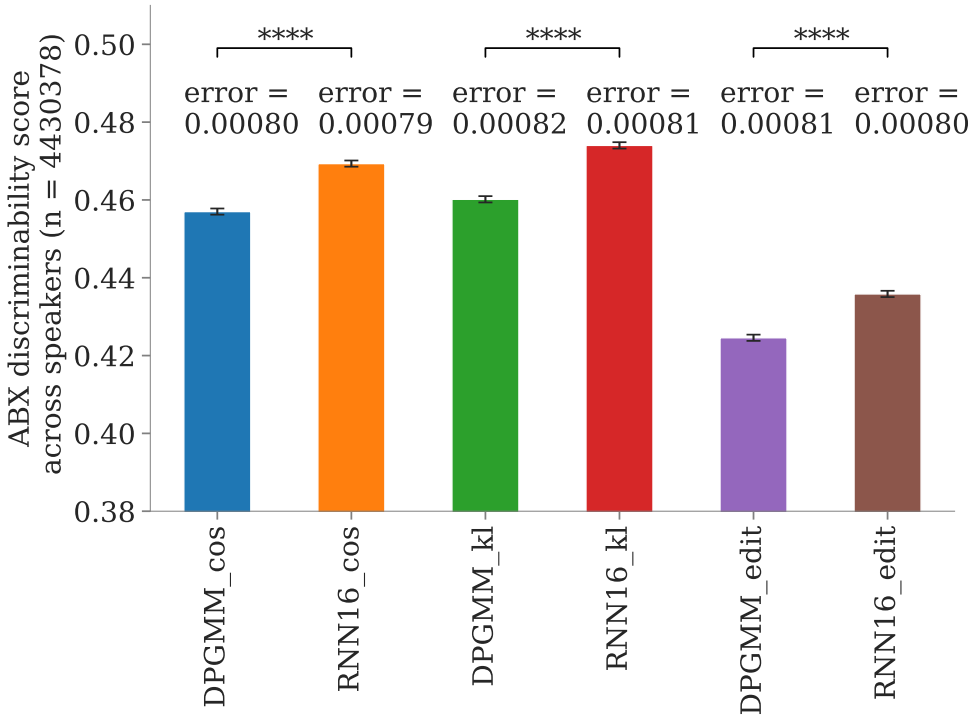


Figure 3.10: Average ABX discriminability score within speakers (upper subfigure) and across speakers (lower subfigure) on TIMIT test corpus. We compared average ABX discriminability scores of all n triplets (A, B and X) between DPGMM algorithm (DPGMM) and DPGMM-RNN hybrid model of 16 contextual frames (RNN16) with cosine distance (cos), Kullback-Leibler divergence (kl), and Levenshtein distance (edit). Significance of paired t-test is indicated by stars: **** means $p \leq 0.0001$, ** means $p \leq 0.01$, and * means $p \leq 0.05$. Error bar is 95% confidence interval; error is annotated above.

3.3.5 DPGMM-RNN Hybrid Model in Zerospeech 2019

Figure 3.11 shows that the DPGMM-RNN hybrid model is better at discriminating phonemes (which is the decrease of the ABX error rate across different distances) and compressing representation (which is the decrease of the bit rate of the one-hot representation of clusters) compared to the DPGMM algorithm.

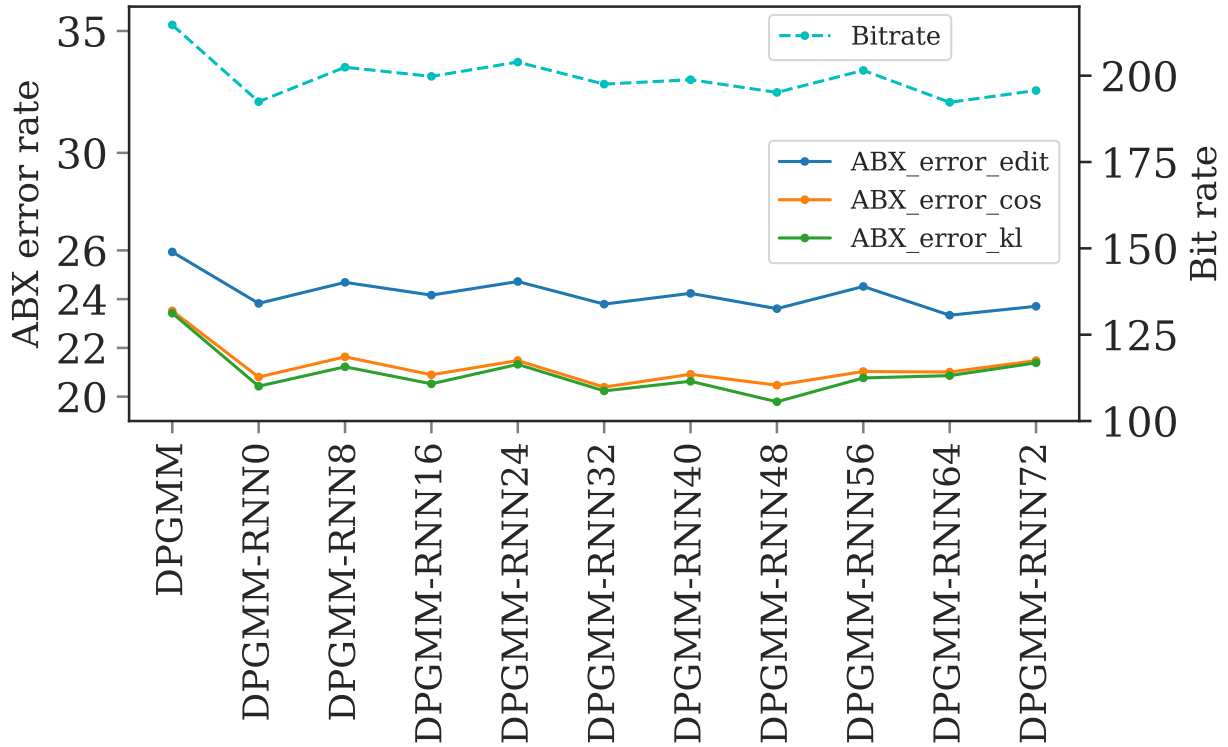


Figure 3.11: ABX error rate and bit rate on English dataset of Zerospeech 2019 [3] (decreases simultaneously with stronger context modeling). Solid lines show ABX error rates with cosine distance, KL divergence and edit distance using the primary vertical axis; dashed line shows bit rate of generated clusters using secondary vertical axis. DPGMM means DPGMM clustering algorithm (without RNN contextual modeling); DPGMM-RNNn (RNNn) means hybrid model (with n RNN contextual frames).

Table 3.2: ABX error rate and bit rate of DPGMM-RNN hybrid models (RNN48 and BiRNN16) and top models from Zerospeech 2019. The provided Zerospeech baseline uses DPGMM clusters trained by variational inference. VQ-VAE extracts discrete representation with speaker-adversarial enhancement (VQ-VAE). Adversarial multi-task learning is used on DPGMM clusters obtained from acoustic features after FHVAE transformation (FHVAE). Our models first get DPGMM clusters (DPGMM) from which we train the DPGMM-RNN hybrid model using the unidirectional RNN with 48 contextual frames (RNN48). Contextual modeling of the hybrid model is further enhanced using the bidirectional RNN with 16 contextual frames (BiRNN16). Numbers of contextual frames of different hybrid models are chosen based on their lowest ABX error rates and lowest bit rates on the Zerospeech dataset.

Method	Baseline	VQ-VAE	FHVAE(a)	FHVAE(b)	DPGMM	RNN48	BiRNN16
ABX_cos	35.63	20.25	13.82	22.32	23.52	20.47	20.08
ABX_kl	34.74	50	13.72	21.67	23.42	19.79	19.97
ABX_edit	35.7	37.31	44.3	26.46	25.94	23.609	22.58
Bitrate	71.98	158.7	1732.81	413	214.69	195.1	188.08

As the number of contextual frames increases, the ABX error rates and the bit rates gradually decrease. We choose RNN48 (about three or four syllables [75] as RNN context) as the result of our DPGMM-RNN model for Zerospeech 2019 because the error rate with the cosine distance and KL divergence start increasing and that with the edit distance is still decreasing.

The Figure 3.11 shows a sharp decrease in the bit rate and the ABX error rates between the DPGMM clustering algorithm and the DPGMM-RNN hybrid model. However, within the DPGMM-RNN models, increasing the length of the contexts slightly decreases the ABX error rate or the bit rate in Zerospeech 2019. The reason might be explained by its English training set, which only contains very short utterances, where the mean duration per utterance is 2.063 seconds, and the three longest utterance durations are 14, 7.99, and 7.82 seconds. When we increase the length of the context of the DPGMM-RNN hybrid model, we expect to capture both the acoustic structure of each phoneme and the statistical structure of a short sequence of several phonemes. This effect of modeling long

contexts is relatively weak because most of the utterances of the English training corpus of Zerospeech 2019 are triphones instead of complete sentences of natural utterances. Longer contextual modeling doesn't show its full power on the dataset when all of the utterances are short.

The VQ-VAE [76, 36] and Factorized Hierarchical Variational Auto-encoder (FHVAE) [77] systems got the top ABX error rate results in Zerospeech 2019 [3]. We compared the system of the DPGMM-RNN hybrid model with these top systems with official toolkits from Zerospeech 2019 (Table 3.2).

The best Zerospeech 2019 system used VQ-VAE [76] to quantize the MFCC acoustic features with several centroids. The system also uses a speaker-adversarial approach [78] to make the final representation speaker independent. Although our system of the DPGMM-RNN hybrid model (RNN48) got a relatively low ABX error rate, it had a slightly higher bit rate than the VQ-VAE based system.

Compared with the DPGMM-RNN hybrid model, the VQ-VAE based system got a much higher ABX error rate with KL divergence because its frame representation was not normalized to be a distribution. The VQ-VAE model got a higher ABX error rate with edit distance because that the DPGMM-RNN hybrid model uses a one-hot vector representation where the maximum edit distance between two frames is 2; the VQ-VAE based system uses discrete representation whose maximum edit distance between two frames might be very large.

Another difference between the two models is that the DPGMM-RNN hybrid model is constrained from accepting Gaussian distributed acoustic features as inputs, and some neural network embeddings containing rich speech information with complex distribution may not work for DPGMM clustering. But VQ-VAE ideally works for any kind of feature.

The second best system first used the FHVAE extracted features to get DPGMM clusters. Those clusters and speaker ids are trained with adversarial multi-task learning to get a final representation. The system has its primary representation (FHVAE(b)) and an alternative (FHVAE(a)) [79].

The system FHVAE(a) gets a very low ABX error rate using continuous representation and high sampling rate to get more acoustic details, which also increases the error rate with edit distance and a very high bit rate. To decrease the bit rate and get discrete representation, the softmax outputs is converted to one-hot rep-

resentations (FHVAE (b)). The system of our proposed DPGMM-RNN hybrid model got a lower ABX error rate across three provided distances and a lower bit rate than the FHVAE (b) system.

The official baseline [39] of Zerospeech 2019 uses compact representation (low bit rate) but sacrifices the discriminability of phonemes (relatively high ABX error rate).

We further enhanced the contextual modeling of the DPGMM-RNN hybrid model using a bidirectional RNN (BiRNN16) as well as the unidirectional RNN (RNN48). Similar to a hybrid model using a unidirectional RNN, the hybrid model using a bidirectional RNN achieved lower ABX error rates (with cosine, KL and edit distances) and a lower bit rate than the DPGMM clustering algorithm. The performance worsened with too many contextual frames because of the limitation of the RNN's contextual modeling ability on the English dataset of Zerospeech 2019 with many short utterances.

The DPGMM-RNN hybrid model using the bidirectional RNN achieved the best performance in the Zerospeech dataset using 16 contextual frames (BiRNN16 with a current frame along with 8 past frames and 8 future frames), which had relatively lower ABX error rates and a lower bit rate than the hybrid system using a unidirectional RNN (RNN48) (Table 3.2).

3.4. Discussion

When we try to identify phonemes from acoustic signals, we directly categorize these units with the non-linear perception from the complex and detailed acoustic signals (bottom-up processing). At the same time, we have abundant high-level knowledge about the statistical structure of phoneme sequences, and this knowledge influences our perception of sound units (top-down processing).

We extracted phoneme-like units using DPGMM clustering on acoustic features, which likes bottom-up processing without high-level knowledge top-down constraints and sometimes concentrates on the local irregularities of the speech, suffering from the fragmentation problem.

We first question whether the fragmentation problem comes from the clustering difficulty of complex acoustical events because people use highly variated

gestures to pronounce sounds with various manners, abstracted as distinctive features.

By exploring the fragmental level of different distinctive features by conditional perplexity (Fig. 3.5), we found that the DPGMM algorithm is worse at categorizing vowels than consonants because it generates more fragmental frames inside the phonemes. A similar situation happens in perception experiments where stably observing the construction of vowel categorization is more difficult than consonant categorization. For example, our perception jumps from one category to another when listening to the stimuli from /p/ to /t/ with equal acoustic changes of the consonants. However, our perception seems more continuous (hearing intra-phonemic variations) than categorical when listening to the stimuli from /i/, /ɪ/ to /ɛ/ with equal acoustic changes of the vowels [57].

The sensitivity of complex acoustic events causes the DPGMM algorithm to suffer from the fragmentation problem. Fig. 3.5 also shows that the fragments of the vowels are associated with the shapes and dynamics of formants. The fragments of the fricatives are associated with the energy concentrated at high frequencies, similar to noise.

Humans can perceive these noisy phonemes with the knowledge of language structure, even when we replace them with actual white noise [59]. This idea inspired us to propose such top-down contextual enhanced methods as RNN and functional load to capture the statistical structure of the acoustical segments for unsupervised phoneme discovery.

Our proposal, the DPGMM-RNN hybrid model, explores how we can use high-level contextual information to relieve the problem of fragmentation. The hybrid model decreases the fragmental level (completeness increase) more than just using DPGMM clustering (Fig. 3.3, Fig. 3.6). Since we experimented on a longer context by taking more frames as RNN input, the fragmental level decreased more. With the same length of contextual frames, considering both the past and future context performances better than just considering one direction (Fig. 3.8).

Enhancing the contextual modeling by RNN helps remove the short-time fragmental segments without generating super-segments that cover several phonemes (Fig. 3.6), as shown by the decrease of the homogeneity and v-measure of our

DPGMM-RNN model (Fig. 3.2, Fig. 3.6). The DPGMM-RNN hybrid model also compressed the segment systems by decreasing the number of clusters (Fig. 3.7).

The DPGMM-RNN hybrid model not only relieves the fragmentation problem but it also finds clusters that more accurately discriminate between phoneme categories. The hybrid model makes less ABX discrimination error (higher discriminability score) and performs more stable (tighter error bar) (Fig. 3.10). The DPGMM-RNN model also got a competitive performance in Zerospeech 2019 in discriminating English triphone segments (Table 3.2).

Clustering algorithms merely based on acoustics, such as the DPGMM clustering algorithm, objectively consider every tiny acoustic detail. The human auditory system has much bias. People are lazy to hear every acoustic detail and merely concentrate on speech units with key information that efficiently conveys the meaning of speech communication (economical principle of speech communication [14]).

The DPGMM-RNN hybrid model uses the RNN to enhance the local contextual model and reduce the fragmentation problem. The DPGMM-RNN model can simulate the human perceptual bias to auto-filling of the noise (fragments) [59], frequent-spoken words [60], and phoneme categorization [57]. The DPGMM-RNN model cannot model the bias induced by the shared cell activations [18] from motor [56] or visual areas [58]. These dynamical activations might rely on the structure determined by gene heredity and spoken and visual habitual experiences.

We reported our results on the same DPGMM setting as the previous works. We also tried other parameter settings [53, 73]. One important one is the concentration parameter that reflects the ability to generate the new clusters. We found that the small concentration parameter always converges but with a slightly slow converge rate. Making concentration parameters large will not decrease the number of clusters. In contrast, the training of DPGMM can be unstable and sometimes never converges. The performance of DPGMM also depends on the distribution of the input features. We attempted some features from neural networks which do not follow the Gaussian distribution, where the DPGMM never converges.

While DPGMM works in the unsupervised phoneme discovery. It is hard to

generate many DPGMM segments long enough to represent the words. Because the model assumes that each feature frame comes from a Gaussian. One Gaussian Cluster might have the limited expression to cover the acoustic variation of a word that contains the variations of several phonemes. Increasing the concentration parameter of DPGMM can slightly make segments longer but not enough to make the segments longer than words.

The DPGMM-RNN hybrid model can generate the DPGMM-RNN labels that are different from the DPGMM labels. These different DPGMM-RNN labels can improve the performance of unsupervised phoneme discovery. That means the wrong predictions by DPGMM-RNN have a positive effect on better clustering of labels. We verified this hypothesis by analysis of the positive correlation between wrong prediction and phoneme accuracy, where the RNN wrong prediction of the DPGMM label is measured by the training loss of the last epoch and the phoneme accuracy is measured by the v-measure.

The frames that RNN labels are different from DPGMM labels mainly come from acoustic-complex phonemes such as fricatives. The DPGMM's framewise prediction suffers from such acoustical complexity. The DPGMM model at frame level also cannot capture the short-time stationary property of speech signals. The DPGMM-RNN hybrid model can capture the short-time stationary property of speech signals by mapping the feature chunks to DPGMM labels.

Chapter 4

DPGMM and DPGMM-RNN Hybrid Model for Low-resource ASR and LVCSR

4.1. Motivation: Unsupervised Empirical Adaptation in Perception Formation Process

Speech feature extraction can affect ASR performance. Such features as Mel-Frequency Cepstrum Coefficients (MFCC) [80] and Perceptual Linear Prediction (PLP) [81] work well in ASR systems using mel-scaled and bark-spaced filterbanks [80, 81] that mimic log-scaled speech perception.

However, speech perception is changed by hearing experiences. Such features as MFCC or PLP, widely used in ASR applications, fail to model the perceptual change due to the past speech learning experiences. Infant perception is changed by listening to speech without text. We propose to model this unsupervised process for feature extraction to improve ASR.

The rest of our introduction is arranged as follows. The first two subsections describe the motivation of our work by arguing that an infant's unsupervised learning experiences change speech perception by causing the permanent brain state modifications that served as a physical fundamental basis for the lifetime speech perception formation process; this realization motivates us to model such

an unsupervised process to improve ASR. The remaining subsections discuss the computational models that are suitable to such an unsupervised learning process of infants in practical and interpretable perspectives and use the features from these models to improve ASR.

4.1.1 Experiences Engraved on Cortex Cells to Affect Perception

Experiences change perception. For example, infants in different countries who are born with similar auditory organs can differentiate phoneme contrasts across languages; their perception is changed to bias their mother tongue after they have more listening experiences [82]. When Japanese infants hear Japanese speech more often from their parents and their surrounding people, they may adapt their perception to become less sensitive to and finally become completely unable to discriminate the phoneme contrast of /l/ and /r/, because this discrimination does not help them differentiate Japanese meanings. In contrast, American infants can discriminate /l/ and /r/ after a year. This empirically adapted perception has long effects in later life as adults.

Empirical adaptation can happen at the organic level. In *On the Origin of Species*, Charles Darwin argued that empirical "habits produce an inherited effect." Here he is relying on his observation of domestic ducks that "the bones of the wing weigh less and the bones of the leg more" compared with wild ducks, because domestic ducks are "flying much less, and walking more" [83]. Experiences can leave "a permanent record ... written or engraved on the irritable substance" [84], and "past occurrences in the history of the organism as part of the causes of the present response" [85]. The term "permanent record" is coined as a "mnemic trace" or an "engram" by the evolutionary biologist Richard Semon [84, 85], who first introduced the concept to the scientific community.

Engram research of mnemic phenomena has recently become an exciting topic in neural science [86]. We intuitively know that if infants play with fire and get badly burned, the painful experience might make them feel fear whenever they see a fire in their lifetime. The key question is whether one can find evidence to support that such experiences actually cause organic changes, especially per-

manent brain changes. Several generations failed for about ten decades until "engram renaissance" [87] started from the early 21th century, sparked by the development of molecular and circuit tools that probe and precisely manipulate brain functions. Neural scientists recently verified the existence of engram cells by tagging the brain cells of mice with stable activations after exposing them to fearful experiences [88]. The tagged cells can be physically manipulated to make mice recall experiences without stimuli [89], disrupt brain records as if such experiences never happened [90], or even implant "fake" memories of non-existing experiences [91, 92]. The endurance of engram changes was verified by measuring the strength of engram cell connections [93, 94].

Neurosurgery studies on patients provide evidence for the neuronal records of engrams. In the Harvey Lecture of 1936, the neurosurgeon Wilder Penfield reported that electrical stimulation on the temporal cortex caused a patient to re-live a frightening childhood episode, which was repeated in her dreams, and she finally freed herself from dream attacks after portions of her right temporal lobe were removed. In the Ferrier Lecture of 1946, Penfield reviewed 190 neurosurgery operations. He determined that stimulation on the temporal lobe created "experiential hallucinations" (the dream-like states) that caused patients to become frightened and cry out. He discovered that stimulation on the temporal lobe created instant "perceptual illusions" that caused patients to alter perceptual interpretations of present experiences [18].

In the early 20th century, in the section of "The Definition of Perception" of the book of *The Analysis of Mind*, Bertrand Russell defined the perception of objects as appearances of objects that "give rise to mnemic phenomena; they are themselves affected by mnemic phenomena" [85]. Russell borrowed the concepts of mnemic phenomena and engrams from Semon. He elaborated the essence of perception in the tradition of Locke [95] and Hume [96], philosophers who in the 17th and 18th centuries argued that such mind-objects as perception come from experiences.

After defining perception in the book, Russell gave the following example that described how current perceptions are affected by past experiences that were engraved in engrams, which are the permanent neuronal records:

> For example, the effect of a spoken sentence upon the hearer depends

> upon whether the hearer knows the language or not, which is a question of past experience... connected with mnemic phenomena...

The engravement of experiences in the brain (the mnemic phenomena that affect perception) of Russell's seminal idea of perception is verified by contemporary neural science that argues that engram cells in the cortex can be 1) activated by learning experiences, 2) physically or chemically modified by learning experiences, and 3) reactivated by subsequent stimuli that represent learning experiences to cause further physical or chemical modifications [86].

In other words, perception starts from experiences and is adapted (changed or affected) by experiences. Speech perception can be adapted by frequent exposure to particular sounds [97]; such adaptations include selective adaptation [98] that lasts for minutes, lexicon adaptation [99] for hours or days (after stimuli of minutes), and language learning adaptation for months or years [17].

4.1.2 Infant Learning Experiences to Establish Lifetime Perception

Speech perception is adapted through language learning experiences (Fig. 4.1). The lifetime speech perception formation process has been initialized at birth. Before exposure to any empirical speech data, such auditory organs as the cochlea are preliminarily sensitive to the range of frequencies within human speech and insensitive to higher frequencies [100].

The primary stage of language learning starts when an infant's "psychological urge" [17] emerges. This urge incentivizes the infants to get what they want or to satisfy a persistent curiosity. They satisfy this desire when they communicate with their parents by unconsciously acquiring spoken language tools and learning to segment and find units inside the speech.

An infant's brain is also "physiologically plastic" [17] for adapting and engraving the neuronal records of word-sounds, concepts, and their connections by frequently listening to the elementary speech from his or her parents that contains fundamental segment units for describing life situations. A neuronal record is formed by the passage of electrical potentials through the nerve cells and over their connecting fibers to alter the states of the engram cells and their nerve

branches and synapses that are waiting to be reactivated or reinforced when similar speech stimuli occur. The formation of such neuronal records allows speech unit retrieval during the process of language learning. Any dysfunction in shaping the neuronal records of speech—the destruction of the "formation of engrams of words" [101]—may cause perception impairment [102], including deafness, aphasia (word-blindness), or agnosia to speech sensory impressions or their association with other mental images. The reinforced engraving of neuronal records can hardly be erased after the first decade of an infant's life; the inevitable decrease of neuronal plasticity increases the difficulty of adding new long-lasting neuronal records in later life [17].

After the primary stage, an infant enters the second stage of language learning called the vocabulary spurt [15] that starts roughly from the second half of the second year. Since toddlers generally can't read or write until about the age of four [17], their speech perception is affected by neuronal records encoding the knowledge accumulated by unsupervised speech learning experiences.

The early infant period of unsupervised empirical adaptation by speech has long-lasting effects in the formation of perception that is further shaped by supervised empirical adaptation when a child eventually learns to write and to build connections between speech and text [15].

Modeling the speech perception formation process (Fig. 4.1) to extract the perceptual features that are related to language learning experiences can improve ASR performance. To model the physiological prior extraction that mimics the log-scaled function of the cochlea, we can extract spectrum features such as MFCC [80] and PLP [81] features; to model the supervised empirical adaptation that learns from speech and text, we can extract supervised features such as bottleneck features (BNF) [13] or language embeddings [103]. However, modeling unsupervised empirical adaptation in the infant period has been less explored for ASR applications (highlighted in red rectangle in Fig. 4.1), especially for Large Vocabulary Continuous Speech Recognition (LVCSR) or low-resource ASR.

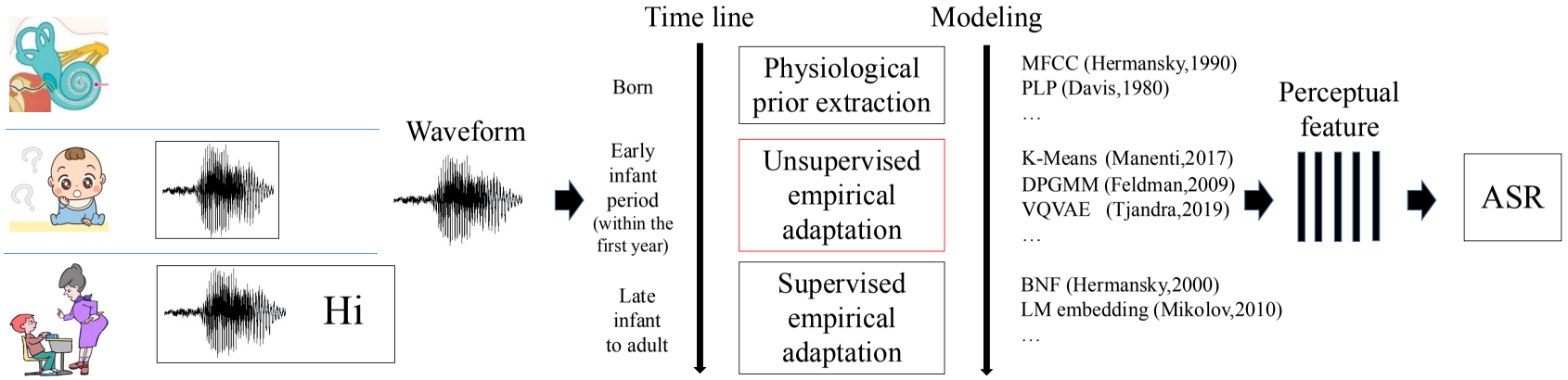


Figure 4.1: Lifetime speech perception formation process. One initializes speech perception by auditory organs after birth, adapts it based on hearing speech without text in the early infant period, and adapts it again by learning experiences that connect speech with text from stages of late infant to adult. Perception shaped in infant period (highlighted by red rectangle) through unsupervised experiences has long-lasting effects on later life. This paper concentrates on utilizing computational models of unsupervised empirical adaptation in infant period to extract perceptual features to improve ASR.

4.1.3 Modeling Unsupervised Empirical Adaptation by DPGMM for ASR

If we believe that speech perception adaptation through experience is an accumulated process from the infant to the adult periods, where each stage might leave organically permanent records, then adaptation in the infant period should have foundational importance in shaping speech perception and language learning. The ASR should improve when we apply the knowledge from the models of unsupervised empirical adaptation of the infant period.

We propose to use the Dirichlet Process Gaussian Mixture Model (DPGMM) [42] to model the unsupervised empirical adaptation to improve ASR for practical and interpretable reasons. DPGMM retained the state-of-the-art performances in the ABX discrimination test at the Zerospeech challenges of 2015, 2017, and 2019 [52, 53, 79]. These Zerospeech challenges aimed to find features strong at identifying and discriminating phonemes from speech without text and compared features that included acoustic features of MFCC or PLP [31], neural network features from autoencoder [32, 33, 34], ABnet [35], and VQ-VAE [36], parametric clustering features from GMM [36] and K-means [36, 37], and nonparametric clustering features from DPGMM trained with Gibbs sampling [52, 53] and variational inference [39, 40]. DPGMM also worked in spoken term detection [38], but it was rarely applied in ASR, especially in LVCSR [104] or low-resource ASR that we will tackle in this paper.

The DPGMM is interpretable as a graphical model [42] that represents conditional dependencies between random variables that 1) show such statistical descriptions as means, variances, and amounts of each potential phoneme-like cluster and 2) show the generative process by unsupervisedly adapting these descriptive parameters to dynamically fit empirical speech data 3) with possible flexible hierarchical extensions [105] that contain more sophisticated explainable linguistic factors, including lexicon or grammar priors [68].

Empowered by its interpretability, in cognitive science, Feldman et al. used DPGMM with a lexicon prior as a computational model to simulate the unsupervised speech learning process of an infant. The simulation illustrated the possible feedback from word segmentation learning that influences phoneme category learning. Such phenomenon challenged and compensated for the traditional view

of the sequential language acquisition of infants from phoneme to word without emphasizing the interaction between the two learning processes [68]. The interactive learning process illustrated by DPGMM was further verified by Feldman et al. to show consistency with infant auditory experiments that demonstrated how word-level information affects the infant perception of phonetic contrasts [28].

This stream of literature aims to use model simulation to illustrate infant distributional learning [106, 107] during phoneme category acquisition and to provide evidence for mechanisms [68, 106] to explain the developmental changes [82, 68, 108] in infant categorical perception. The related research used computational models of unimodal, bimodal, GMM, and DPGMM with rich information from the descriptive statistics of modals (simulating linguistic categories) and flexible extensions to integrate more knowledge such as lexicons. Maye et al. used the unimodal or bimodal frequency distribution [106] to demonstrate an infant's sensitivity to the statistical distribution of speech sounds. Boerl and Kuhl et al. used GMM with the EM algorithm [107] to illustrate that infants can learn more easily and accurately with infant-directed speech than adult speech. McMurray et al. used a GMM with gradient descent [108] to introduce the continuous development trajectories of the infant distributional learning of phoneme categories. Feldman et al. used a non-parametric Bayesian approach of DPGMM to study feedback mechanisms from word learning to phoneme learning [68]. Feldman's finding of the interactive learning process of the infants using DPGMM is well referenced by cognitive science, psychology, and infant language acquisition.

4.1.4 Modeling Unsupervised Empirical Adaptation by DPGMM-RNN Hybrid Models for ASR

However, DPGMM fails to model the temporal order of speech features [63], because the Dirichlet Process (DP) of DPGMM is theoretically infinitely exchangeable, meaning that the joint distribution of DPGMM does not depend on the order of data if they are infinite [64]. The weak framewise temporal modeling increases the model sensitivity to local trivial random acoustic details. Such sensitivity makes DPGMM clustering uncertain for assigning clusters to frames (Fig. 4.2) and creates small, random cluster segments inside a phoneme (Fig. 2.3).

This is DPGMM's "fragmentation problem" [4].

In unsupervised phoneme discovery, DPGMM tends to suffers from a fragmentation problem when the model encounters the frames from such acoustically complex phonemes as a fricative with noise-like high frequencies or a vowel with rapid formant transitions [4, 1]. DPGMM tends to generate more clusters than the number of phonemes in any human language [53, 1] when it struggles to discriminate between complex phonemes with higher resolution.

We propose to use the DPGMM-RNN hybrid model [4], which enhances DPGMM, to model unsupervised empirical adaptation to improve ASR. The DPGMM-RNN hybrid model 1) improves temporal modeling and 2) relieves fragmentation problems of DPGMM with RNN to relearn the connection between acoustic features and DPGMM cluster labels or posterior vectors by listening to feature chunks instead of concentrating on trivial details at the frame level like DPGMM.

In unsupervised phoneme discovery, the DPGMM-RNN hybrid model enhances temporal modeling to improve its capturing of such important acoustic cues as the formant transitions that occur within diphthongs, the coarticulation effects from adjacent phonemes, and the suprasegmental factors over phonemes. The DPGMM-RNN hybrid model relieved the fragmentation problem and decreased the fragmental level measured by the conditional perplexity [109] and the v-measure [72]. It also reduced the number of clusters of DPGMM [4] and overperformed DPGMM in unsupervised phoneme discovery on datasets from Zerospeech 2019 with an ABX discrimination test at a moderate bitrate [4].

Inspired by the relation between engram and perception, we use DPGMM and DPGMM-RNN hybrid model to extract perceptual features. The engrams that encode past speech experiences can transform sensations into perception, where Russell [85] defined the sensations as the parts inside perception without influence from the past experiences. For example, by retrieving the language knowledge from the learning experiences that are stored at the engram, we transform our sensation of the sound to our perception of the speech. Our computational model parameters that encode past empirical speech data (after adapting parameters to fit the data) can transform the present sensational features into the perceptual features, where the sensational features include MFCC that has a log-scale audi-

tory property.

In summary, we propose to use DPGMM and the DPGMM-RNN hybrid model to model the unsupervised empirical adaptation and extract perceptual features to improve ASR (Fig. 4.1), where these perceptual features extend MFCC features with DPGMM or DPGMM-RNN posteriorgrams by concatenation (Fig. 2.4).

The rest of this article is arranged as follows:

1. We verify the effectiveness of our proposed features with the ASR system on the English corpora of TIMIT [2] and WSJ [12] (a widely used dataset for LVCSR) and on the low-resource corpora of Mboshi [6] and Javanese [7] (a telephone conversation dataset that roughly contains a three-hour training set with hundreds of speakers from different dialect regions talking under noisy environments).

2. We compare the ASR performance of unsupervised DPGMM features from our proposal with the supervised bottleneck features (BNF) from Kaldi [5].

3. In the discussion section, we scrutinize that the DPGMM and DPGMM-RNN model perplexities agree with infant perceptual perplexity from auditory experiments. Our analysis provides evidence to support our hypothesis that our proposed features reflect unsupervised perception adaptation at an early infant period.

4. In the discussion section, we examine relation between the perception formation process and the ASR results.

4.2. Experiments

4.2.1 Datasets and Their Divisions

TIMIT

We analyzed the models on the TIMIT corpus [2] of English read speech because it includes reliable and detailed phoneme annotations. We followed the official division [2] of a training set of 3.14 hours, a development set, and a complete test set of 1344 utterances.

WSJ

We checked the LVCSR performance on the WSJ corpus [12] of the English speech. We followed the official division [12] of the training datasets of WSJ SI-84 of 15.08 hours and WSJ SI-284 of 81.25 hours, an identical development dataset called dev93, and an identical evaluation dataset called eval92.

Mboshi

We further experimented on a low-resource African read corpus of Mboshi [6] that is spoken in Congo Brazzaville and Diaspora. It has a writing system developed by missionaries without standardized orthography. The Mboshi text mainly comes from the Bible. The corpus extracted all the spoken sentences from a Mboshi-French dictionary [110] and a fieldwork-oriented Bouquiaux and Thomas's corpus [111].

The Mboshi dataset [7] officially contains training and development sets. We divided the original training set into a development set of 200 utterances and a training set of remaining utterances and treated the original development set as a test set. The development set took the first few utterances (Table 4.1) of each speaker with the roughly same ratio of utterances per speaker in the original training dataset that contains sorted utterances according to utterance ids. We computed the durations after trimming the head and tail silences (Section 4.2.2). Table 4.1 summarizes the statistics of the Mboshi dataset.

Javanese

We attempted some challenging experiments on a low-resource Indonesia telephone conversational corpus of Javanese [7] that represents its Central, Western, and Eastern dialect regions. These telephone calls were recorded by hundreds of speakers from 16 to 65 years old of roughly equal genders using different models of mobile phones (e.g., Nokia, Sony) by different networks (e.g., Smartfren, XL) or using landlines in various environments, including cars, offices, streets, and public places.

We divided the Javanese dataset based on the utterance order in demographics.tsv, which is a documentation file that accompanied the data release [7] that

Table 4.1: Statistics of low-resource Mboshi read speech datasets [6] of three speakers.

Mboshi	**#Hours**	**#Utterences**	**#Utterences/speaker**
Train	2	4416	3186 / 1060 / 170
Development	0.07	200	144 / 48 / 8
Test	0.21	514	351 / 126 / 37

contains the information of the utterances grouped by speakers, in the following steps:

- The dataset with 6720 utterances was decreased to 3749 utterances after removing those that contained tokens of <X>, including <non-speech> and <int> (interrupt), and it was further decreased to 3157 utterances after removing the utterances that only contained one token.
- We then divided the 3157 utterances with the first 200 utterances as a development set, the second 200 utterances as a test set, and the remaining 2757 utterances as a training set.
- To ensure that the divisions contained no speaker overlap, we adjusted the 217, 194, and 2746 utterances as development, test, and training sets by the utterance order of the records (grouped by speakers) in demographics.tsv.
- To ensure that no text overlap exists in the division between the test set and the training or development sets, we removed the utterances from the test set whose texts occurred in the training or development sets. Finally, we got 217, 155, and 2746 utterances as development, test, and training sets for our experiments.

We computed the durations after trimming the head and tail silences (Section 4.2.2). Table 4.2 summarizes the statistics of the Javanese dataset.

Table 4.2: Statistics of low-resource Javanese telephone datasets [7]. The 3-hour conversational dataset was recorded by hundreds of speakers from different dialect regions using different mobile devices under various noisy backgrounds, where the designed division was non-overlapping in speakers or sentences between test set and training set or development set.

Javanese	**#Hours**	**#Utterences**	**#Speakers**	**#Speakers/gender**
Train	2.88	2746	201	F 100 M 101
Development	0.2	217	14	F 8 M 6
Test	0.17	155	15	F 6 M 9

4.2.2 Feature Extraction

Acoustic feature extraction

We followed Kaldi [5] using a 39-dimensional MFCC+Δ+$\Delta\Delta$ (25-ms frame size and 10-ms frame shift) with mean and variance normalization (CMVN) as the acoustic feature for TIMIT and a 40-dimensional MFCC of high resolution with CMVN as the acoustic feature for WSJ. We used the identical feature setup as TIMIT for the Mboshi and Javanese corpora that have similar data amount as TIMIT.

VAD for low-resource corpora

We found utterances in Mboshi and Javanese have long head and tail silences (sometimes over five seconds), with which our encoder-decoder attentional ASR struggled. We did energy-based Voice Activity Detection (VAD) for both corpora.

For the Mboshi corpus, since we found that the officially provided alignments of silences from a light-weight ASR toolkit [112] failed to precisely perform VAD, we trimmed the head and tail silence segments whose maximum absolute amplitudes were smaller than the threshold of 0.1.

For the Javanese corpus, the VAD with a fixed amplitude threshold failed because the complex recording devices and environments made utterances whose sounds were weaker than the noisy silences of other utterances. We dealt with the problem by a simple method called dynamical VAD that halved the initial

threshold of 0.1 several times until the trimmed audio had more than 100 samples for each utterance.

DPGMM and RNN posteriorgram extraction

We extracted the DPGMM posteriorgrams with a basic implementation that strictly followed the steps described in the method section without any optimizations or approximations. In our practice, we found a simple implementation with Numpy without GPU optimization, with several hundred lines of codes, provided an acceptable speed for our experiments.

Instead of independently applying the DPGMM algorithm on the test set, we froze the DPGMM parameters adapted by the training sets and used these fixed parameters to generate DPGMM posteriorgrams for the development and test sets.

The training process for DPGMM used the same parameter setup as previous works [53, 4, 73]. We set the concentration parameter to 1, the mean and variance of the priors to the global mean and global variance of the MFCC features, and the belief-strengths of the mean and the variance to 1 and $D + 2$, where D is the dimension of MFCC. We obtained clusters and posteriorgrams after 1500 sampling iterations.

We extracted RNN posteriorgrams from the DPGMM-RNN hybrid model [4] and fed the RNN with the MFCC feature chunk of a center frame binding with eight left and eight right adjacent frames. We used an RNN of a 5-layer BiLSTM whose input layer size matched the MFCC dimension, whose output layer size matched the number of DPGMM clusters, and whose hidden layer size was 512. The training of RNN used 20 epochs with a batch size of 256.

Table 4.3: Hyperparameters for encoder-decoder ASR and DPGMM. Notion D is number of dimensions of MFCC features.

Model	Parameters	Value
ASR	Dropout probability	0.25
	Label-smoothing ratio	0.05
	Learning rate	0.001
	Beam size	10
DPGMM	Concentration parameter	1
	Belief-strength of mean	1
	Belief-strength of variance	$D + 2$
	Belief of mean	Feature mean
	Belief of variance	Feature variance
	Number of iterations	1500

4.2.3 Attentional Encoder-Decoder ASR System

We used pytorch to implement an ASR system of an attentional encoder-decoder model [8] that consisted of a three-layer pyramid bidirectional LSTM encoder [8] that had 256 hidden units at each direction and dropped half of the frames to reduce the time resolution by a factor of 2 at each layer, a decoder [9] that contains a single-layer LSTM with 512 hidden units, and MLP attention [9].

MLP attention scheme generated the expected contextual vector by a probability vector output from a fully connected layer (MLP) fed with the concatenation of the current decoder hidden state and the encoder output (contextual vector). Table 4.4 shows that the decoder [9], at each time step, was fed with the concatenated feature of the output from the embedding layer and output from the previous decoding step, which was further processed by the LSTM and dropout layers. The output that was concatenated with the expected contextual vector from the attention was fed into a fully connected layer of 256 hidden units, followed by a tanh activation function. For the encoder with three layers, we dropped half of frames at each layer such that the 3-layer encoder output has a length that is 1/8 of the number of frames of the current utterance features,

Table 4.4: Architecture of attentional encoder-decoder ASR system. A → B denotes next layer of layer A is layer B. pBiLSTM denotes a pyramid bidirectional LSTM [8]; FC stands for a full-connected layer; EMBED denotes an embedding layer. Module-N denotes module with N hidden units (e.g., FC-512 denotes a fully connected layer with 512 hidden units). Contextual FC-256 is a fully connected layer fed with current embedding concatenated with expected contextual vector from attention. At each time step, the decoder, proposed by Luong [9], is fed with a concatenated feature of output of decoder pre-net and output of decoder from previous step. Encoder input is acoustic features; input of decoder pre-net is characters. pBiLSTM uses dropout regularization at each layer.

Module	**Cascaded layers of module**
Encoder	FC-512 → ReLU → Dropout → 3-layer pBiLSTM-256 (reduce half of the frames per layer)
Decoder pre-net	EMBED-256 → Dropout
Decoder [9]	(Pre-net output + Prev. decoder output) Single-layer LSTM-512 → Dropout → Contextual FC-256 → Tanh
Decoder post-net	Softmax
MLP attention	FC-256 → Tanh

which decreased the number of frames and captured the contexts across successive frames.

In the encoding stage, we fed speech features into a fully connected layer of 512 hidden units, followed by a ReLU activation function and a dropout layer with probability 0.25 before the pyramid BiLSTM. On the decoding stage, we put each character into an embedding layer of 256 hidden units, followed by a dropout layer before the decoder, whose output was converted into a probability vector by a softmax layer. For the MLP attention, we used one hidden layer of 256 units, followed by a tanh activation function. We used weight tying [113]

between the input and output embeddings and label smoothing [114] with a ratio 0.05 in the decoder. We used the weight normalization in the attention.

When we trained the ASR system, we set the batch size to 32 and used the Adam optimizer [115] with an initial learning rate of 0.001, which decreased by a half whenever the loss successively increased for more than three epochs. Our ASR systems usually converged between 30 and 70 epochs after the learning rate dropped below 1e-5. We used a gradient norm clipping strategy [116] when training each batch to deal with the problems of exploding and vanishing gradients.

We evaluated our ASR system with a beam search where the beam size was 10 and the expand size [117] (which denotes as the maximum candidates per node to introduce more diversity into the search) was 5. We also increased the penalty [118] for long sentences with coefficient 1.

All reported ASR results in the paper are from this ASR system without any pronunciation dictionaries or language models in the decoding process.

In summary, we used an attentional encoder-decoder ASR system [8] which includes an encoder of a three-layer LSTM, an attention of a Multi-Layer Perceptron (MLP), and a decoder of a one-layer LSTM. The setups of the ASR include dropout probability as 0.05, label smoothing ratio as 0.05, learning rate as 0.001 (which halves whenever the training loss successively increases for more than three epochs), and beamsize as 10.

The ASR and DPGMM hyperparameters are summarized in Table 4.3 and the structure of our attentional encoder-decoder ASR system is summarized in Table 4.4.

4.3. Results

4.3.1 Discriminative Posteriorgram and Fragmentation Problem

We concatenated the MFCC features with the posteriorgrams from the DPGMM clustering algorithm or the DPGMM-RNN hybrid model. We describe the characteristics of these posteriorgrams using an utterance from the TIMIT test set.

Fig. 4.2 shows that the DPGMM posteriorgram discovered those phonemes with stable acoustics (see the red rectangles). However, it suffers from fragmentation problems from complex acoustics (see the black circles). The fragmentation problems represent the uncertainty of the DPGMM algorithm when judging the cluster assignment to each frame.

Fig. 4.2 also shows that the RNN posteriorgram (from the DPGMM-RNN hybrid model) can relieve the fragmental problems from the DPGMM posteriorgram on such phonemes with complex acoustics as fricatives that contain noise-like high-frequency components (see the black circles).

4.3.2 Fragmentation Problem and ASR Error

We analyzed the potential relations between the fragmentation characteristics and the ASR performance of the proposed features. We measured the ASR performance by counting the ASR phoneme errors of the TIMIT test set by comparing the annotated references with the recognized hypotheses; the references and hypotheses were aligned to have the same length by sclite [5] for each utterance. We analyzed the decrease of the phoneme errors by the categories of distinctive features (rather than deletion, insertion, and substitution categories). For example, the number of phoneme errors of the distinctive features of the stops is the number of stop consonants in the test set whose ASR alignments mismatch the underground annotations; the decrease of the phoneme errors of the stops from the MFCC features to their concatenation with the DPGMM posteriorgrams (MFCC_vs_MFCC+DPGMM in Fig. 4.3) is the difference of the number of phoneme errors of the stops before and after concatenation, which indicates an ASR improvement of the proposed feature compared to the MFCC feature.

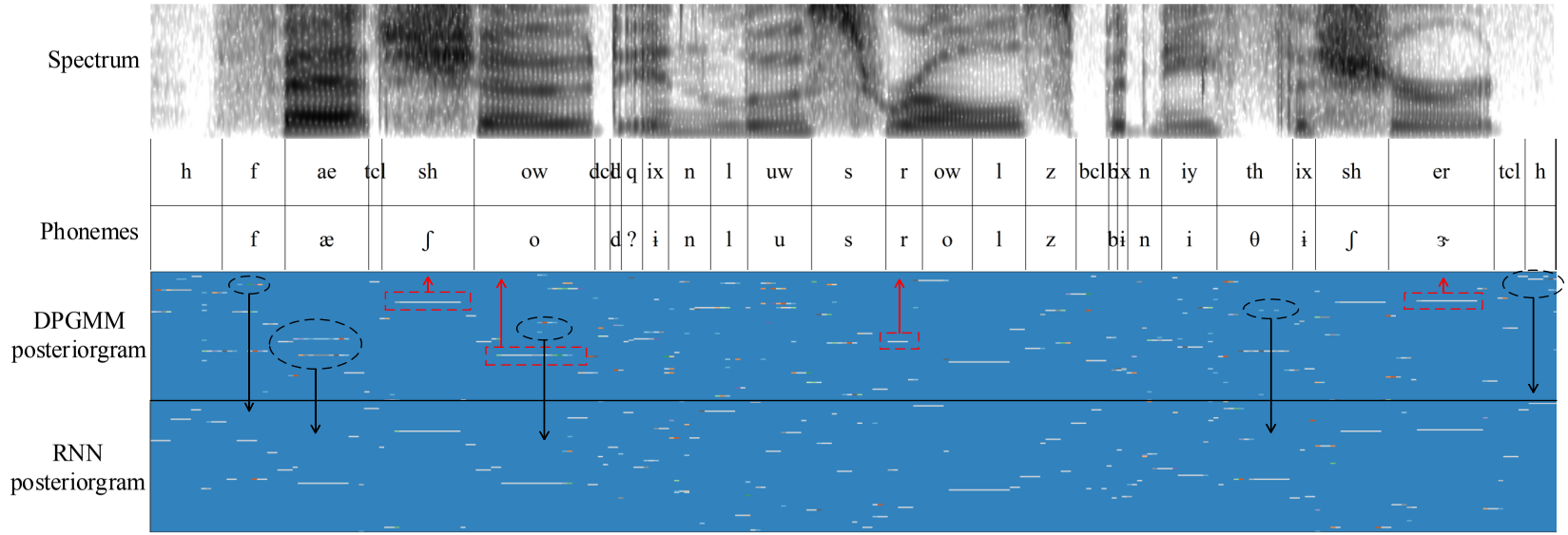
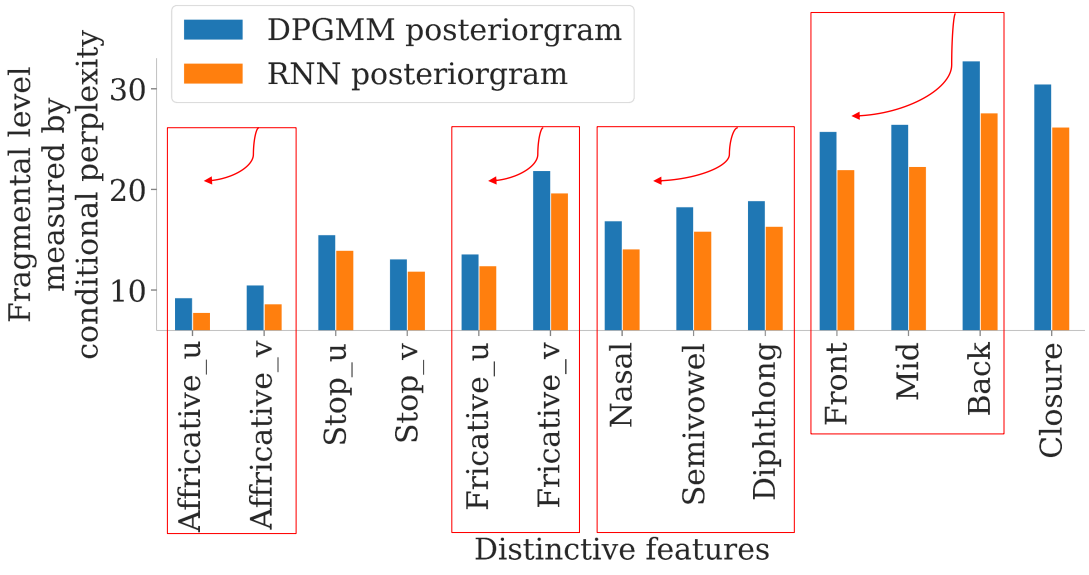


Figure 4.2: Phoneme recognition (red rectangles) and fragmentation problem (black circles) of posteriorgrams. Utterance "Fat showed in loose rolls beneath the shirt" with id FADG0_SI1909 from TIMIT test set shows posteriorgrams from DPGMM clustering algorithm (DPGMM posteriorgram) and DPGMM-RNN hybrid model [4] (RNN posteriorgram). Top layer is spectrogram followed by phoneme layer, DPGMM, and RNN posteriorgram layers. Red rectangles show DPGMM posteriorgram discovered phoneme segments to improve phoneme recognition; black circles show RNN posteriorgram relieved fragmental problem (uncertainty in cluster assignment) of DPGMM posteriorgram.

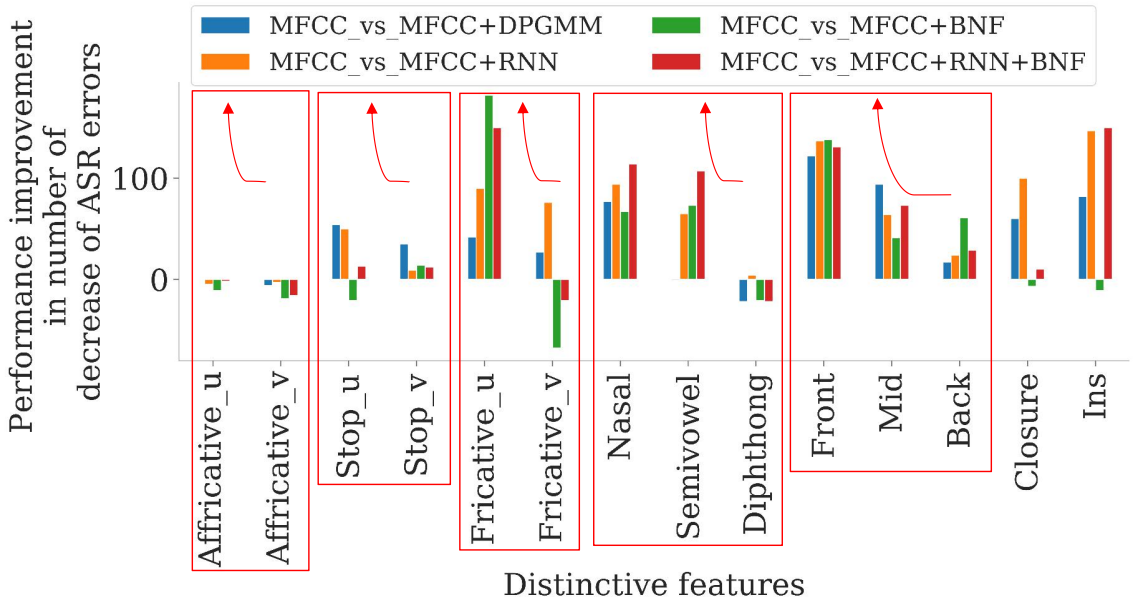
The ASR improvement, indicated by the decreased number of ASR errors, is induced by the proposed feature extension with DPGMM or RNN posteriorgrams characterized by the severity of their fragmentation problems. We measured the fragmental level of the posteriorgrams by the conditional perplexity of the clusters given phonemes [4], which is the exponential of conditional entropy [109] that reflects the average number of DPGMM or RNN clusters per phoneme.

Feature extensions with the posteriorgrams of different fragmental levels change the phoneme error distribution of the ASR system. Fig. 4.3 shows the following results.

- Unvoiced consonants, less fragmental than voiced ones, tend to have more ASR improvement.
- Vowels from back to front that are less fragmental tend to have more ASR improvement when their first and second formants become less compacted and easier to differentiate.



(a) The fragmental level measured by conditional entropy



(b) The ASR improvement measured by error decrease

Figure 4.3: Fragmental levels and ASR improvements of distinctive features on TIMIT test set. Upper subfigure (a): conditional perplexity of cluster given phonemes [4] that shows fragmental level of posteriorgrams from DGPMM algorithm (DPGMM posteriorgram) and DPGMM-RNN hybrid model [4] (RNN posteriorgram) for each distinctive feature. Lower subfigure (b): decrease number of phoneme errors that shows ASR improvements from MFCC acoustic features to their concatenations with DPGMM posteriorgrams (MFCC_vs_MFCC+DPGMM) and from MFCC features to their concatenations with RNN posteriorgram (MFCC_vs_MFCC+RNN) for each distinctive feature; we also added results of bottleneck features (BNF) from Kaldi default scripts. Red rectangles with arrows show tendency between decrease of fragmental level and improvement of ASR performance among distinctive features. stop_v denotes voiced stop; stop_u denotes unvoiced stop. Ins denotes insertion errors of ASR that inserts symbols not in reference phonemes. Closure includes silences and short pauses.

- The RNN posteriorgram relieves the fragmental problem of the DPGMM posteriorgram [4] (Fig. 4.2), indicated by decrease of fragmental level mea-

sured by conditional perplexity for each distinctive feature (Fig. 4.3a). The concatenation of the MFCC feature with the RNN posteriorgram (MFCC+RNN) tends to achieve more ASR improvement than concatenation with the DPGMM posteriorgram (MFCC+DPGMM) (Fig. 4.3b).

- The MFCC feature extension with the RNN posteriorgram (MFCC+RNN), compared with the DPGMM posteriorgram (MFCC+DPGMM), tends to have more ASR improvement on such complex acoustics as fricatives containing noisy, high-frequency components, diphthongs with complex formant structures, or closures with various silences (sometimes with background noises) and short pauses (Figs. 4.3b and 4.2).

- Unsupervised DPGMM based features (MFCC+DPGMM and MFCC+RNN) work well at silences (closure). The RNN context enhancements (MFCC+RNN and MFCC+RNN+BNF) help remove insertion errors. The unsupervised features compensate for the supervised features (MFCC+BNF vs. MFCC+RNN+BNF) in ASR.

4.3.3 Evaluation by Large Vocabulary Continuous ASR

Our preliminary analysis on the TIMIT corpus show that our proposed feature extension improved the simple ASR of read speech. The improvement on the simple ASR drove us to explore the performance of our proposed features on a more challenging LVCSR task on the WSJ corpus of the WSJ SI-284 set (an 80-hour training set) and the WSJ SI-84 set (a 15-hour training set).

We first attempted to directly feed the DPGMM or RNN posteriorgrams into the ASR system because the DPGMM posteriorgrams effectively discriminated the phonemes on several Zerospeech challenges [29, 30, 3], and the DPGMM-RNN hybrid model outperformed the DPGMM clustering algorithm at discriminating and identifying phonemes [4]. Table 4.5 shows that the RNN posteriorgram (RNN) worked better than the DPGMM posteriorgram (DPGMM) in ASR, but neither reached the ASR performance of MFCC.

Table 4.5: LVCSR performance on WSJ. We compared MFCC features, DPGMM posteriorgrams, RNN posteriorgrams, and their concatenations on our attentional encoder-decoder ASR system, along with two baselines [10, 11], by character error rates (CERs) on WSJ speech corpus [12], including training datasets of WSJ SI-84 that is about 15 hours or WSJ SI-284 that is about 80 hours, without pronunciation dictionaries or language models in decoding process. Both baselines used Mel-scale filterbank coefficients (MEL) that are frequency-domain equivalent forms of MFCC features. The WERs were consistent with the CERs. On our encoder-decoder ASR without a language model, our proposed feature concatenation achieved a 15.25% in WER on WSJ SI-284 set, compared with a previous report of 18.2% [10].

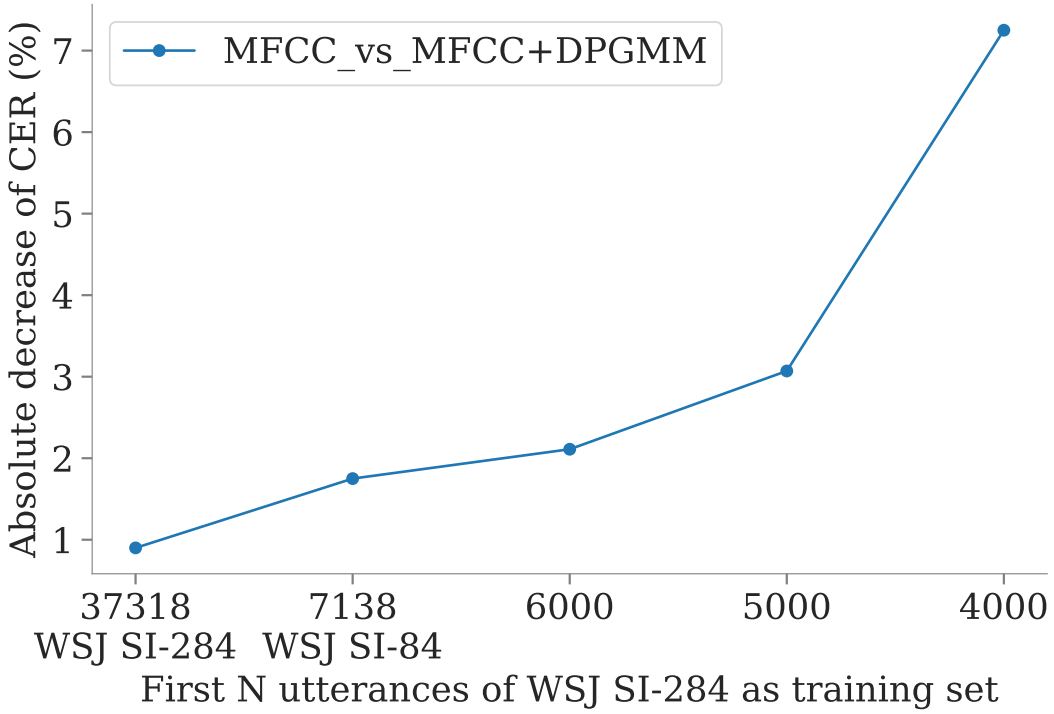
System with feature	**WSJ SI-84 CER(%)**	**WSJ SI-284 CER(%)**
Baseline ASR1 MEL [10]	17.01	8.17
Baseline ASR2 MEL [11]	17.35	7.12
Our ASR MFCC	16.61	6.57
Our ASR DPGMM	35.5	12.35
Our ASR RNN	29.21	11.27
Our ASR MFCC+DPGMM	14.86	5.67
Our ASR MFCC+RNN	14.25	5.55

We further attempt to concatenate the MFCC features with the DPGMM or RNN posteriorgrams. Although the posteriorgrams strengthened the discrimination capability on acoustically stable phonemes, they suffer from fragmentation problems on acoustically complex phonemes (Fig. 4.2) that can be compensated by MFCC features. Table 4.5 shows that the concatenated features (MFCC+DPGMM or MFCC+RNN) got fewer ASR errors than the MFCC features (MFCC); the concatenated features converged faster and retained the improvement of the character accuracy of the development set during the training process better than the MFCC features [104].

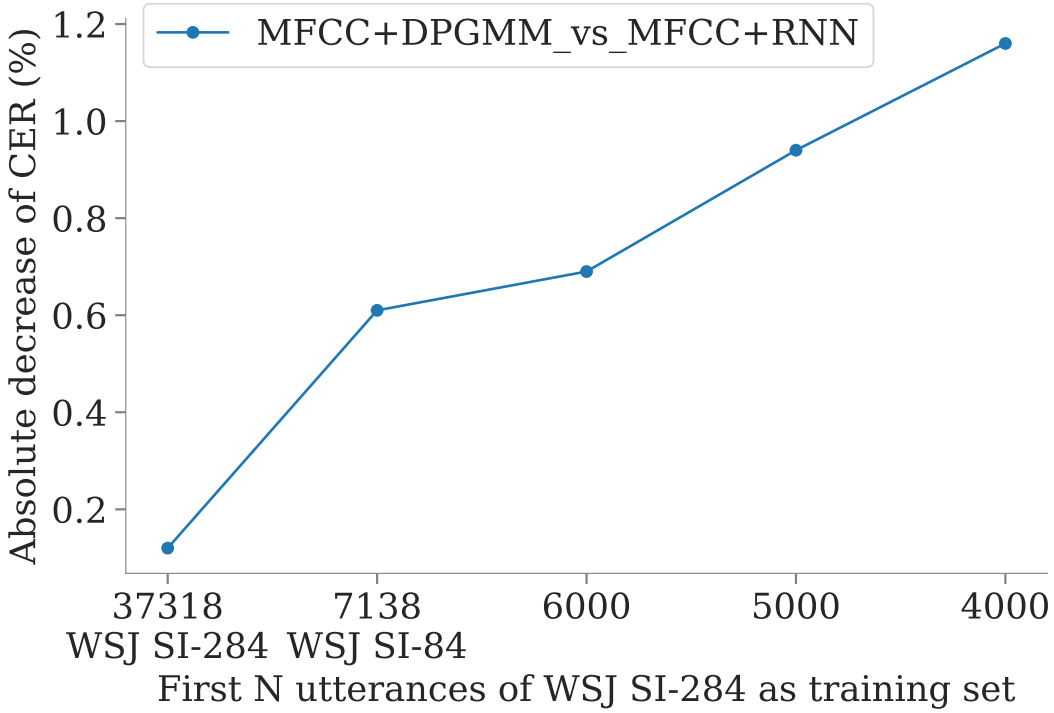
Table 4.5 shows that the feature extension with the RNN posteriorgram (MFCC+RNN) achieved a lower CER than that with the DPGMM posterior-

gram (MFCC+DPGMM); both of the proposed feature extensions outperformed the MFCC feature (MFCC). The WERs were consistent with CERs; our proposed feature extension (MFCC+RNN) achieved a WER of 15.3%, compared with 18.2% in a previous work [10], on the WSJ SI-284 set with an encoder-decoder ASR without a language model.

We observed that the absolute ASR improvement, from the MFCC feature (MFCC) to its DPGMM feature extension (MFCC+DPGMM), on the WSJ SI-284 set is smaller than that on the WSJ SI-84 set (0.9% and 1.75% respectively in Table 4.5). We explored the relation between the absolute ASR improvement and the amount of data. We trained the ASR system by the first N utterances of the WSJ SI-284 training set to examine the change of the absolute ASR improvement when N became smaller, until the data amount was too small to support ASR (Fig. 4.4).



(a) ASR improvement from MFCC to MFCC+DPGMM



(b) ASR improvement from MFCC+DPGMM to MFCC+RNN

Figure 4.4: ASR tendency with less data. Upper subfigure: ASR improvement from MFCC feature to concatenation of MFCC feature and DPGMM posteriorgram (MFCC_vs_MFCC+DPGMM). Lower subfigure: ASR improvement from DPGMM posteriorgram (MFCC+DPGMM) to RNN posteriorgram (MFCC+RNN). We trained ASR with the first N utterances of WSJ SI-284 set, where the first 37318 utterances are the WSJ SI-284 set and the first 7138 utterances are the WSJ SI-84 set. The CERs of ASR trained with first 3000 utterances exceed 80% (not shown in figures) and that of first 4000 utterances were about 40%.

Fig. 4.4a shows that extending the MFCC feature with the DPGMM posteriorgram (MFCC_vs_MFCC+DPGMM) improved the ASR performance more with less data. Fig. 4.4b shows that enhancing the DPGMM posteriorgram with the RNN posteriorgram (MFCC+DPGMM_vs_MFCC+RNN) improved the ASR performance more with less data.

4.3.4 Evaluation by Low-resource Read and Telephone ASR

Our LVCSR results on WSJ show that the proposed feature extensions are more effective with less data. This finding suggests a potential of our proposed features for a low-resource ASR when low-resource languages lack a well-studied written form with limited speech data that have annotations transcribed by expert linguists mainly from fieldwork (e.g., Mboshi) or when the low-resource languages have limited annotated data (e.g., Javanese). We verified the effectiveness of our proposed features on the low-resource ASR.

Table 4.6: ASR performance on low-resource corpora. We compared MFCC features (MFCC) and their feature extensions with DPGMM and RNN posteriorgrams (MFCC+DPGMM and MFCC+RNN) by ASR error rates on low-resource speech corpora of Mboshi [6] and Javanese [7] and on TIMIT [2] as a simulation of a low-resource corpus due to its small data amount. Feature extraction and ASR system of three corpora shared identical scripts with identical parameter setups.

Feature	**Javanese CER(%)**	**Mboshi PER(%)**	**TIMIT PER(%)**
MFCC	53.23	22.67	23.92
MFCC+DPGMM	51.68	20.91	22.74
MFCC+RNN	48.19	20.67	22.38

We treated TIMIT as a simulation of a low-resource dataset because it has a small amount of data close to the other two low-resource datasets. The Mboshi read speech dataset has been well recorded, annotated, and checked by linguists. The Mboshi is officially divided into the training and development sets that contain three overlapped speakers. Table 4.6 shows that the ASR on Mboshi outperformed TIMIT.

The Javanese telephone conversation dataset included hundreds of speakers whose ages ranged from 16 to 65 using 24 types of mobile devices or landlines through eight types of networks in seven types of environments with speech representing three different dialect regions. Some utterances were weak and hard to

hear clearly; some were recorded under loud background noises; the annotation of the Javanese dataset was relatively difficult and noisy. The dataset division did not overlap between speakers or sentences. Table 4.6 shows lower ASR performance on Javanese than TIMIT or Mboshi.

Table 4.6 shows that the feature extensions by the DPGMM or RNN posteriorgrams (MFCC+DPGMM or MFCC+RNN) had better ASR performances than the MFCC features (MFCC).

Table 4.6 further shows that the RNN posteriorgram extension (MFCC+RNN) improved over the DPGMM posteriorgram extension (MFCC+DPGMM) and more improvement on Javanese than on Mboshi and TIMIT. The noisy Javanese corpus made DPGMM relatively unstable. The RNN posteriorgrams with RNN contextual enhancement stabilized the DPGMM posteriorgrams and made them more robust on noisy Javanese compared to Mboshi and TIMIT.

We compared the DPGMM-RNN feature with the VQCPC feature in the low-resource ASR of Javanese telephone speech. We extracted the VQCPC feature from the state-of-art implementation in the Zerospeech 2020 [119]. The DPGMM-RNN features performed better than the VQCPC feature on the noisy low-resource ASR (CER 48.19% vs. 50.12%). The VQCPC feature from its autoencoder module might suffer from over-rich representation risky in learning the noise. In contrast, DPGMM-RNN uses the GMM to fit the Gaussian-distributed MFCC feature that may be more robust to noise. The DPGMM-RNN model also has the RNN to enhance contextual modeling.

4.3.5 Comparsion and Combination with Supervised BNF in ASR

Both supervised BNF [13] and unsupervised DPGMM-RNN features [4] help increase the ability of acoustic features to discriminate phonemes. It would be more persuasive to show the effectiveness of our proposed unsupervised DPGMM-RNN features by comparison with the widely-used supervised BNF features with a reliable implementation.

The BNF feature needs accurate alignments (estimated starting and ending times of each phoneme) to work well; Kaldi [5] is state-of-art for this purpose.

Table 4.7: ASR performance of unsupervised and supervised features. We compared unsupervised feature extension with RNN posteriorgrams [4] (MFCC+RNN) with supervised feature extension with BNF features [13] (MFCC+BNF). For WSJ and TIMIT, we used Kaldi's [5] official scripts without modification for ASR alignment and BNF extraction; for Javanese and Mboshi, we followed the Kaldi scripts of TIMIT. The following table includes ASR results of the concatenated features by MFCC, RNN, and BNF (MFCC+RNN+BNF). The abbreviations of the recording devices of TEL, MOB, and MIC denote telephones, mobiles, and microphones. The WSJ corpus contains spontaneous dictation from journalists.

Feature	**Javanese CER(%)**	**Mboshi PER(%)**	**TIMIT PER(%)**	**WSJ SI-284 CER(%)**
Data amount (hours)	2.88	2.00	3.14	81.25
Data style	Spontaneous	Read	Read	Read
Recording devices	TEL/MOB	MIC	MIC	MIC
MFCC	53.23	22.67	23.92	6.57
MFCC+RNN	48.19	**20.67**	**22.38**	5.55
MFCC+BNF [5]	47.47	21.93	23.04	5.00
MFCC+RNN+BNF	**43.23**	21.26	22.53	**4.48**

For example, Kaldi attained a WER of 2.3% [120] on the WSJ corpus with a hybrid system and a language model; the performance is a breakthrough achievement for existing ASR implementations. We obtained the ASR alignments for BNF extraction on WSJ with the official Kaldi scripts without modification. We extracted the BNF features of WSJ and TIMIT with the official Kaldi script for BNF extraction (run_bnf.sh) without modification except for changing paths to datasets; we believe the default settings of the BNF script were well tested and tuned. We obtained the alignments and extracted the BNF of Mboshi and Javanese following the Kaldi scripts of TIMIT because these datasets have similar data amounts.

The Kaldi toolkit extracted the BNF features by training a 5-hidden-layer neural network with 1024 hidden dimensions and 42 bottleneck dimensions to map each frame of the MFCC feature concatenated with four left frames and four right frames to alignments generated by a system pipeline of monophone training, triphone training, LDA transformation, MLLT transformation, and speaker adaptive training (SAT) [5].

Compared with the BNF features with dense representations (similar values in every dimension) whose segment boundaries are affected by the given ASR alignments, the DPGMM or DPGMM-RNN posteriorgrams with sparse representations (compressing information in a few dimensions) have phoneme discriminability affected by the fitness of the MFCC acoustic distributions to Gaussian mixture assumptions. The sparseness of the posteriorgrams removes the redundancies for phoneme discrimination between acoustic stable segments; the overcompression with information loss of the posteriorgrams causes instabilities for segment judgment on acoustic complex phonemes, such as noisy fricatives.

In other words, the alignment-based BNF features and the Gaussian-based DPGMM-RNN features capture different discrimination information dependent on the supervised ASR alignments and the unsupervised Gaussian fitness. The two types of features improve different perspectives of the ASR and can compensate for each other. Table 4.7 shows that ASR achieved better performance on the concatenation of MFCC, RNN, and BNF (MFCC+RNN+BNF) than the concatenation of MFCC and BNF (MFCC+BNF) for all the corpora.

Table 4.7 shows that the combination of MFCC, RNN, and BNF features

(MFCC+RNN+BNF) worked best on WSJ and Javanese. In both cases, supervised BNF features (MFCC+BNF) overperformed unsupervised DPGMM-RNN features (MFCC+RNN). Because the BNF features (MFCC+BNF) from a neural network were supervised by huge-data-guided reliable ASR alignments on WSJ and were supervised by noise-resistant annotated training text on Javanese compared with DPGMM-RNN features (MFCC+RNN) that are sensitive to noise and have no text supervision.

Table 4.7 also shows that a combination of MFCC and RNN posteriorgram (MFCC+RNN) achieved the best performance for Mboshi and TIMIT. In both datasets, the small amount of data, just several hours, caused difficulties in learning reliable alignments and training a neural network to extract BNF. The data condition of the read speech of Mboshi and TIMIT is clean enough to reflect the Gaussian-distributed nature of the MFCC features to extract reliable DPGMM-RNN features.

We build models for incremental perceptual learning at different stages. We model the newborn infant's innate auditory perception with the MFCC feature whose filterbank module simulates the cochlea's nonlinear auditory sensation. We model the young infant's unsupervised speech perception with the DPGMM that simulates the infant's unsupervised speech parsing and classification practice before knowing the written language. We model the grow-up infant's supervised speech perception with BNF that simulates the infant linguistic feature learning by associating the speech signals with linguistic units after the infant learns the textual language and acquires basic word-phoneme correspondence knowledge. The whole incremental perceptual learning process to recognize the speech can be simulated by a combination of the MFCC, DPGMM, and BNF features.

4.4. Discussion

4.4.1 Linking DPGMM Computational Perplexity, Infant Perceptual Perplexity, and ASR error

One slippery, fundamental question is whether such computational features as DPGMM (or DPGMM-RNN) features can be called 'perceptual' and can match

human categorical perception, especially infant perception that is both not fully developed [16] and different from adult perception [82]. That is, can we show evidence that DPGMM categorizes speech well where infants perceive well and that DPGMM categorizes speech poorly where infants perceive poorly. Our DPGMM analysis by conditional perplexity on TIMIT (Fig. 4.5) shed light on this question.

We define the DPGMM perplexity of phonemes as conditional perplexity of DPGMM clusters given the phonemes [4] where conditional perplexity is the exponential of the conditional entropy [109]); we define the DPGMM perplexity of a distinctive feature as the DPGMM perplexity of phonemes with that distinctive feature.

Our analysis on the conditional perplexity on TIMIT (Fig. 4.5) shows the following associations between DPGMM computational perplexity and infant perceptual perplexity on phonemes. The DPGMM (or DPGMM-RNN) perplexity of the consonant stops is relatively low among all the distinctive features. There exists extensive literature about infant perception of stops. Eimas et al. found that 1- and 4-month-old infants can perceptually categorize the stop consonants (/b/ and /p/) [121]. Bertoncini et al. further found that 4- to 5-day-old neonates can discriminate the stops of consonants /b, d, g/ in an environment of a vowel /a/ or /i/ [122]. Stops are among the easiest and the earliest distinctive features perceived by infants.

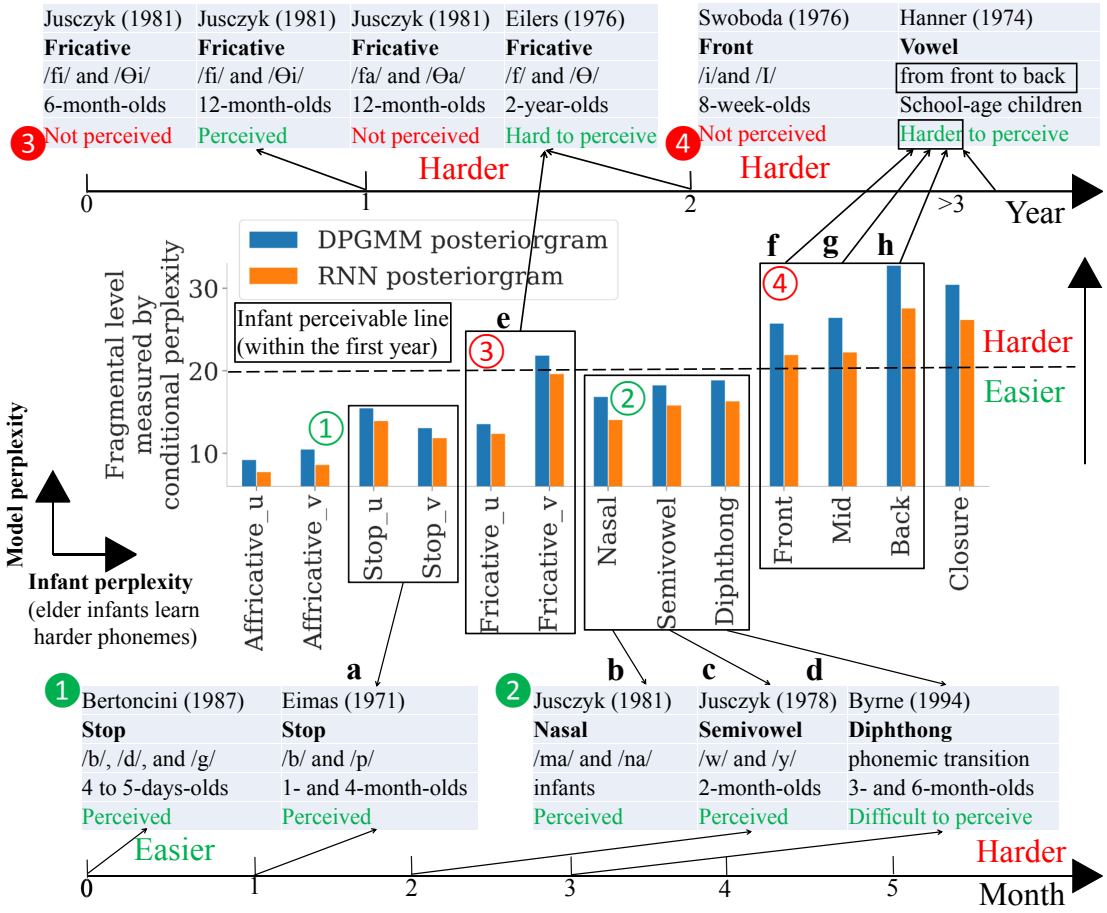


Figure 4.5: Relation between DPGMM model perplexity on TIMIT corpus and infant perceptual perplexity by auditory experiments. Circled numbers denote degrees of perplexity, including DPGMM and DPGMM-RNN model perplexity vertically and infant perceptual perplexity horizontally. Infant perceivable line divides distinctive features that are easy (green) or hard (red) for infants to discriminate.

The DPGMM perplexity of vowels is higher than consonants, and voiced consonants are higher than unvoiced ones. Trehub et al. examined infant vowel discrimination (/i/ vs. /u/ and /a/ vs. /i/) but could not determine whether infants can discriminate vowels categorically, as they did for stop consonants [123].

Their work inspired Swoboda et al. to start the very first systematic study, and they found that 8-week-olds discriminate vowels (/i/ vs. /I/) in a continuous as opposed to a categorical manner [124].

The DPGMM perplexity of fricatives is high among the consonants. Fricatives /f/ and /θ/ are fragmental with high perplexity and are frequently observed in individual utterance examples (Fig. 4.2) of DPGMM clustering [4]. Eimas et al. found that 6- and 12-month-olds cannot discriminate /fa/ and /θa/; only 12-month-olds can discriminate /fi/ and /θi/ [125]. The contrast of /f/ and /θ/ is difficult for toddlers as well. Eilers and Oiler reported on 2-year-olds [126]; Abbs and Minifie reported on preschool children from 3- to 5-year-olds [127].

The DPGMM perplexity becomes higher from front vowels to back vowels. Swoboda et al. showed that 8-week-olds cannot categorize front vowels [124]. The accurate discrimination of vowels by school-age children, in the phonemic environment of /r/, is ranked roughly from front to back [128].

The DPGMM-RNN perplexity is smaller than the DPGMM perplexity in semivowels, diphthongs, and nasals, because DPGMM does not involve temporal order modeling [64], and the DPGMM-RNN hybrid model involves temporal order modeling that may help capture such important temporal cues as formant transitions. Jusczyk et al. found that 2-month-olds discriminated semivowels (/w/ and /y/) based on formant transition differences [129]. Byrne et al. found that 3- and 6-month-olds can discriminate interphonemeic transitions inside a diphthong [130]. Nasals (/ma/ and /na/) can be distinguished by formant transitions [125, 131].

Our further analysis (Fig. 4.3b) suggests a potential causal relation between DPGMM computational perplexity and DPGMM ASR performance because our proposed feature extension with DPGMM posteriorgrams changed the ASR error distribution. Fig. 4.3b shows that several groups, including unvoiced vs. voiced consonants, indicated by red rectangles, of related distinctive features with lower perceptual perplexity tended to have more ASR improvement.

4.4.2 Modeling Perception Formation Process for ASR with Exposure to Different Data Amounts and Data Complexity

Our results agree with our intuition under the perception formation process (Fig. 4.1). The DPGMM or RNN posteriorgrams (modeling unsupervised empirical adaptation) compensate for the MFCC features (modeling physiological prior) and the BNF features (modeling supervised empirical adaptation). Table 4.7 shows that in all corpora, feature extension with RNN posteriorgram (MFCC+RNN) works better than MFCC feature (MFCC); feature extension with RNN and BNF features (MFCC+RNN+BNF) outperformed the feature extension with BNF features (MFCC+BNF).

When exposed to a small amount of clean data in TIMIT and Mboshi, the unsupervised features (MFCC+RNN) worked better than the supervised features (MFCC+BNF) (Table 4.7). This result implies that unsupervised adaptation hugely impacts the early perception-shaping process of the infant period under limited speech exposure in a relatively stable environment with simple and limited linguistic knowledge. Both TIMIT and Mboshi are well designed and professionally annotated with fair details and high accuracy through several rounds of careful checks for delicate analysis in phonetic and linguistic research. Our experiments on WSJ also support that unsupervised adaptation modeling shows more power when exposed to less speech data (Fig. 4.4).

When exposed to a large amount of data in WSJ or complex noisy data in Javanese, unsupervised features compensated the supervised features (MFCC+RNN+BNF vs. MFCC+BNF) to improve the ASR performance (Table 4.7). Supervised adaptation with a neural network (MFCC+BNF) that simulates a late period of supervised acquisition, such as powerful learning and robust discrimination of adults, greatly improved the recognition performance (Table 4.7). Unsupervised adaptation has quite different discriminative power compared to supervised adaptation and serves as a basis for future supervised adaptation.

Computational models encode experiences as parameters. For example, the neural networks including feedforward and recurrent neural networks parameterize the empirical data using continuous and dense representations with enormous

connections. The graphical models including GMM and DPGMM parameterize the empirical data using probabilistic relations among a few key factors. We can use their empirically adapted encoding that compresses the information from history speech stimuli to transform the raw sensational features of the current speech stimuli to an approximation of empirically affected perceptual representations. For example, we used DPGMM training parameters to transform the MFCC features into perceptual features. These computational encodings are likely to be associated with biological experimental measurements of engrams. The construction mechanism of the sparse engram complex revealed by neural science are likely to provide a hint for how to computationally render efficient and informative experiential encoding. It would be interesting to explore, from biological and computational perspectives, the mechanism of how engrams (from experience to encoding) affect the transformation (from sensation to perception). The future exploration that advances these research problems will deepen our knowledge about the perception formation mechanisms and inspire the construction of novel and practical learning algorithms.

Chapter 5

Conclusion and Future works

5.1. Conclusion

In this thesis, we have two proposals. For the first proposal, to mimic the human perception bias of phonemes over acoustic signals, we proposed the DPGMM-RNN hybrid model to improve phoneme categorization. Results show that with the DPGMM-RNN hybrid model, we can relieve the fragmental problem and improve phoneme discriminability

Our second proposal used the DPGMM algorithm and the DPGMM-RNN hybrid model to model the unsupervised empirical adaptation to extract perceptual features to improve ASR. We found that our proposed unsupervised DPGMM and DPGMM-RNN features achieved better performance than MFCC features on the LVCSR and the low-resource conversational ASR.

We compared our proposed unsupervised DPGMM-RNN features with the supervised bottleneck features from Kaldi; the ASR results demonstrate that 1) unsupervised features outperformed supervised features on small and clean datasets; 2) unsupervised features compensated for the supervised features on huge or noisy data datasets.

Our analysis on TIMIT that discloses the relation between the DPGMM computational perplexity and the infant perceptual perplexity provides evidence to support our declaration that the proposed features reflect the infant perception, whose phonemic categorizations are not fully developed.

The analysis on TIMIT also supports our arguments that 1) the DPGMM

and DPGMM-RNN hybrid model with adapted parameters that encode empirical speech data, same as the engrams that encode the knowledge learned from the experience of hearing speech, can transform sensational features into perceptual features; 2) we can improve the ASR performance using the perceptual features of our proposed DPGMM or DPGMM-RNN features compared to the sensational features of MFCC that fail to model the influence from the past experiences.

Our results of unsupervised phoneme discovery and low-resource ASR show that the DPGMM-RNN feature is better than the DPGMM feature; the DPGMM feature is better than the MFCC feature. The MFCC and PLP features use windows and filterbanks to summarize the spectrum context of a few successive acoustic samples. Such resolution is relatively fine and involves noises. The DPGMM on MFCC and PLP features start another level of the contextual model, but a weak one, on the DPGMM and PLP frames. Such DPGMM roughly remembers the frequency or unigram of the cluster history as the experiences or contexts. The DPGMM-RNN model uses RNN to learn wilder context from DPGMM features and captures higher-level context. Such multi-step modeling of hierarchical contexts might generate a more stable representation at each step from MFCC, DPGMM, to DPGMM-RNN features.

Our analysis of the unsupervised phoneme discovery improvement measured by v-measure from DPGMM feature to RNN feature supports the temporal enhancement from the DPGMM and the DPGMM-RNN model. Our analysis of ASR improvement (on TIMIT) shows that the obvious insertion error decrease indicates 1) the highest ASR improvement from MFCC feature to DPGMM-RNN feature and 2) the highest ASR improvement from DPGMM feature to DPGMM-RNN feature among the ASR improvements by other error decreases of different auditory features. We explain that the incremental contextual model from the MFCC feature to the DPGMM-RNN feature has stabilized the local acoustic representations that smoothed the irregularities of MFCC feature inducing the ASR insertion of random labels. Such incremental contextual model decreases insertion error from MFCC to DPGMM based features.

5.2. Future works

Our two proposals explored to find the universal methods at the feature level to improve acoustic modeling of ASR at different supervised levels that the Zerospeech and ASR communities have been using different methods to deal with. In the future, we would attempt to extend this work to universal methods at the syntactic and semantic levels to find hierarchical structures or semantic embeddings to improve unsupervised phoneme discovery, low-resource ASR, and LVCSR tasks.

We proposed the universal acoustic models to find speech units as a human does in speech recognition at the different supervised levels including only speech as infants, a small amount of parallel data of speech and text as children, and a large amount of parallel data as adults. In the spoken language learning, we learn from limited samples of parallel speech and text that contain the connection between speech and text from parents, schools, and dictionaries. We learn from abundant samples of speech when we communicate with different talkers (without text but with larger acoustic variation). We learn from abundant samples of text when we read different books (without speech but with complex semantic contexts). From a practical view, a weak ASR has a potential to become stronger if it can learn from a large amount of audio or text files on the internet. In future work, we also want to extend our framework to include and handle such cases as with a mixture of a small amount of parallel data and a large amount of unparallel speech (or text).

Our proposed DPGMM-RNN model uses the DPGMM module to find the spectral modes and the RNN module to smooth the local irregularities. The DPGMM sometimes maps the different phonemes with similar acoustics into the same cluster type. For example, some vowels such as nasals are voiced and mapped to the same DPGMM cluster. The study of vowel sounds shows stable discrimination by the formants. If the vowels only differ slightly by formants, such a small difference might be hard to detect by the DPGMM clustering algorithm. A decision tree that asks the formant-related questions on the DPGMM frame might obtain higher formant discriminations for vowels.

The DPGMM fragmentation problem is relieved by the RNN but far from being solved. The conditional entropy analysis shows that most phonemes cor-

respond to more than five DPGMM-RNN clusters. We also observe that many DPGMM-RNN segments capture only portions of a phoneme. One potential solution might be to introduce the hidden Markov states into the DPGMM model to capture the temporal information, which might be risky to increase the model complexity and increase the training cost. The Markov state may be transient or recurrent depending on the distribution of speech data. Another potential solution is the multi-resolution DPGMM-RNN hybrid model. The DGPMM assumes that a frame comes from a Gaussian. Then we can induce that the concatenation of several frames is also a Gaussian. Under such insight, we can concatenate several frames of a segment into a new frame to generate the low-resolution DPGMM labels. Such labels learn segments rather than frames and may ignore more frame-wised acoustic details that might cause fragments. The RNN can learn both from low-resolution DPGMM labels and original DPGMM labels to learn the hierarchal hidden structure and may further relieve the fragmental problems.

Several future directions about model structure optimization deserve more exploration. First, such VQVAE based models as VQCPC uses self-supervised learning techniques to learn the hidden representations. The model optimization of such hidden representation can help in downstream tasks of speech technology. The interpretation of the representation of VQVAE or VQCPC needs some exploration. Second, we want to enhance the syntactic and semantic modeling for the DPGMM-RNN feature, where the syntactic and semantic processing are widely studied in NLP. The transformer enhancement is promising due to its success in NLP tasks. Thus we can integrate the powerful NLP models such as a transformer or at least combine the self-attention modules with the RNN modules of the current proposed model. One core challenge would be analysis about improvements of the phonological, grammatical, or semantic linguistic structures that underlie the speech due to transformer enhancement. Third, it would be meaningful and challenging to design reasonable casual assumptions and add such factors to enrich the Baysian network of DPGMM.

In this thesis, we scrutinized the perception bias in phoneme discovery. We proposed the DPGMM-RNN hybrid model to deal with several failures in DPGMM's modeling of perceptual bias. We analyzed how the DPGMM-RNN model achieves better phoneme categorization (a typical perceptual bias). We also emphasized

speech perception formation by the language learning experience. We proposed the DPGMM-RNN feature to improve the low-resource ASR. We provided evidence about the relation between the DPGMM based features and infant unsupervised perceptual adaptation. We want to continue our exploration of the fundamental research programs of this thesis: the mechanism of infant perception for phoneme discovery and the formation of speech perception by the learning experiences. Now we list some potential research directions for further explorations.

The first future direction is studying the mechanism under speech perception by the brain activities using deep learning. 1) First, apply deep unsupervised learning to obtain hidden representations that approximate the underlying distribution of the high-dimensional source data from speech signals, brain activations, and motor activities to enhance the speech-related downstream tasks. 2) Second, design information-rich and interpretable causal models including graphical models to simulate the distributions of signals from speech, brain, and motor activities and induce the relationships between the brain and motor activities and acoustic signal of the speech. 3) Third, use deep learning to learn the distribution of the functional representation in the brain and dig its relation to categorical perception.

The second future direction is modeling the perceptual learning process from experiences. 1) First, we want to find representations of neural records of speech experiences. We can use deep learning technology to find better representations from speech or brain images. For example, a direct application of DPGMM to brain images to see the possible distribution. 2) Second, we can use reinforcement learning to study the speech learning that is associated with environments such as infant-directed speech, infant self-exploration activities, or parent's cultivation activities. We need to design and abstract the language learning rewards of infant interaction with environments of real life.

The third future direction is to study the evolutional view of categorical perceptual instinct determined by genes. 1) First, study how speech perceptual formation from the genetic view. The process of the raw perception to the mature perception of an infant of periods, from a cell to the early-age infant, might reflect the evolution from cell to animals to human, a highly selection-survived process that forms instincts as gene expressions. In such a view, the gene is a

recorder of the successful survival experiences from ancestors. In the genes, some important experiences are encoded and expressed as instincts including categorization of speech, a basic language ability of infants. 2) Second, explore the problem of how the instinct of auditory or speech parsing is possible. Speech perception, as an auditory ability, might have some transitional periods from an audition of sound to the perception of language. Such transition period might remain in some animals with some rudimentary process of language or speech. For such animals that have categorical perception of sound, one problem is that is there any gene expressions that determine such categorization instinct as the evolutionary evidence to human language ability. 3) Third, exploration of the genetic root of speech perception and empirical formation of speech perception through early infant activity or animal experiments also might provide a new perspective to deal with language disorders.

The fourth future direction is to study the cellular root of language perceptual learning ability. 1) First, explore the possibility of using brain imaging or calcium imaging to observe brain activities to explore the association between the functional areas and the perception behaviors. 2) Second, explore the possibility of using optogenetics to manually turn on or off a few target neurons in these areas to observe the causal relation between neuron cell firing and perceptual behavior for determining the 'language' neurons. 3) Third, revisit speech learning about these 'language' neuronal plasticity that reflects language learning ability. 4) Fourth, explore the possibility of using iPS cell technology to recover the plasticity of 'language neuron'.

References

[1] Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura. Optimizing DPGMM clustering in zero-resource setting based on functional load. In *SLTU*, volume 1, pages 1–5, 2018.

[2] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech Disc 1-1.1. Technical Report 93, 1993.

[3] Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W Black, et al. The Zero Resource Speech Challenge 2019: TTS without T. In *INTERSPEECH*, pages 1088–1092, 2019.

[4] Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura. Tackling perception bias in unsupervised phoneme discovery using DPGMM-RNN hybrid model and functional load. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:348–362, 2021.

[5] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The Kaldi speech recognition toolkit. In *ASRU*, pages 1–4, 2011.

[6] Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, H'el'ene Maynard, Markus Müller, Annie Rialland, Sebastian Stüker, François Yvon, and Marcely Zanon Boito. A very low resource language speech corpus for computational language documentation experiments. *A Computing Research Repository (CoRR)*, abs/1710.03501, 2017.

[7] Aric Bills, Thomas Conners, Anne David, Luanne Dela Cruz, Eyal Dubinski, Jonathan G. Fiscus, Ketty Gann, Mary Harper, Michael Kazi, Hanh Le, Nicolas Malyska, Jennifer Melot, Jessica Ray, Fred Richardson, Anton Rytting, and Jacqui Zwanenburg. IARPA Babel Javanese Language

Pack IARPA-babel402b-v1.0b LDC2020S07. *Linguistic Data Consortium (LDC)*, 2020.

[8] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In *ICASSP*, pages 4960–4964, 2016.

[9] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421, 2015.

[10] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *ICASSP*, pages 4835–4839, 2017.

[11] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989, 2020.

[12] Douglas B Paul and Janet Baker. The design for the wall street journal-based CSR corpus. In *ICSLP*, pages 899–902, 1992.

[13] Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *ICASSP*, pages 1635–1638, 2000.

[14] Martinet André. Economie des changements phonétiques. *Francke*, 1955.

[15] Suzanne Curtin, Daniel Hufnagle, Karen Mulak, and Paola Escudero. *Speech perception: development*, pages 1–7. Elsevier, 2017.

[16] Rebecca E Eilers, Wesley R Wilson, and John M Moore. Developmental changes in speech discrimination in infants. *Journal of Speech and Hearing Research*, 20(4):766–780, 1977.

[17] Wilder Penfield and Lamar Roberts. *Speech and brain mechanisms*. Princeton University Press, 1959.

[18] Wilder Graves Penfield. Ferrier lecture — some observations on the cerebral cortex of man. *the Royal Society of London. Series B-Biological Sciences*, 134(876):329–347, 1947.

[19] Amit Das, Jinyu Li, Changliang Liu, and Yifan Gong. Universal acoustic modeling using neural mixture models. In *ICASSP*, pages 5681–5685. IEEE, 2019.

[20] Hui Lin, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, and Chin-Hui Lee. A study on multilingual acoustic modeling for large vocabulary asr. In *ICASSP*, pages 4333–4336. IEEE, 2009.

[21] Jeremy Reed and Chin-Hui Lee. A study on music genre classification based on universal acoustic models. In *ISMIR*, pages 89–94, 2006.

[22] Matthias Boesing, Andreas Hofmann, and Rik De Doncker. Universal acoustic modelling framework for electrical drives. *IET Power Electronics*, 8(5):693–699, 2015.

[23] Lawrence Rabiner. Fundamentals of speech recognition. *Fundamentals of Speech Recognition*, 1993.

[24] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29, 2012.

[25] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006.

[26] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent NN: First results. *arXiv preprint arXiv:1412.1602*, 2014.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[28] Naomi H Feldman, Emily B Myers, Katherine S White, Thomas L Griffiths, and James L Morgan. Word-level information influences phonetic learning in adults and infants. *Cognition*, 127(3):427–438, 2013.

[29] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. The Zero Resource Speech Challenge 2015. In *INTERSPEECH*, pages 3169–3173, 2015.

[30] Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. The Zero Resource Speech Challenge 2017. In *ASRU*, pages 323–330, 2017.

[31] Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH*, pages 1781–1785., 2013.

[32] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater. A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge. In *INTERSPEECH*, pages 3199–3203, 2015.

[33] Leonardo Badino, Claudia Canevari, Luciano Fadiga, and Giorgio Metta. An Auto-encoder based approach to unsupervised learning of subword units. In *ICASSP*, pages 7634–7638, 2014.

[34] Leonardo Badino, Alessio Mereta, and Lorenzo Rosasco. Discovering discrete subword units with binarized Autoencoders and Hidden-Markov-Model encoders. In *INTERSPEECH*, pages 3174–3178, 2015.

[35] Roland Thiolliere, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *INTERSPEECH*, pages 3179–3183, 2015.

[36] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. VQVAE unsupervised unit discovery and

multi-scale Code2Spec inverter for Zerospeech Challenge 2019. In *INTERSPEECH*, pages 1118–1122, 2019.

[37] Céline Manenti, Thomas Pellegrini, and Julien Pinquier. Unsupervised speech unit discovery using K-means and neural networks. In *International Conference on Statistical Language and Speech Processing (SLSP)*, pages 169–180, 2017.

[38] Chia-ying Lee and James Glass. A nonparametric Bayesian approach to acoustic model discovery. In *ACL*, pages 40–49, 2012.

[39] Lucas Ondel, Lukáš Burget, and Jan Černocký. Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86, 2016.

[40] Janek Ebbers, Jahn Heymann, Lukas Drude, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. Hidden Markov model variational Autoencoder for acoustic unit discovery. In *INTERSPEECH*, pages 488–492, 2017.

[41] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[42] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

[43] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[44] Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *Technical Report, University of British Columbia*, 1, 2007.

[45] Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[46] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[47] Thomas Schatz. ABX-discriminability measures and applications. *Ph.D. Thesis*, 2016.

[48] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. Unsupervised learning of acoustic sub-word units. In *ACL*, pages 165–168, 2008.

[49] Marijn Huijbregts, Mitchell McLaren, and David Van Leeuwen. Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *ICASSP*, pages 4436–4439, 2011.

[50] Aren Jansen and Benjamin Van Durme. Efficient spoken term discovery using randomized algorithms. In *ASRU*, pages 401–406, 2011.

[51] Alex S Park and James R Glass. Unsupervised pattern discovery in speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, 2008.

[52] Michael Heck, Sakriani Sakti, and Satoshi Nakamura. Feature optimized DPGMM clustering for unsupervised subword modeling: a contribution to Zerospeech 2017. In *ASRU*, pages 740–746, 2017.

[53] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li. Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: a feasibility study. In *INTERSPEECH*, pages 3189–3193, 2015.

[54] Gordon E Peterson and Harold L Barney. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184, 1952.

[55] Leigh Lisker and Arthur S Abramson. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422, 1964.

[56] Charles Francis Hockett. *A manual of phonology*. Waverly Press, 1955.

[57] Alvin M Liberman, Franklin S Cooper, Donald P Shankweiler, and Michael Studdert-Kennedy. Perception of the speech code. *Psychological review*, 74(6):431, 1967.

[58] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

[59] Richard M Warren. Perceptual restoration of missing speech sounds. *Science*, 167(3917):392–393, 1970.

[60] William F Ganong. Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, 6(1):110, 1980.

[61] Mark A Pitt and James M McQueen. Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, 39(3):347–370, 1998.

[62] Jason Chang and John W Fisher III. Parallel sampling of DP mixture models using sub-cluster splits. In *NIPS*, pages 620–628, 2013.

[63] Dilan Görür and Carl Edward Rasmussen. Dirichlet process Gaussian mixture models: choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

[64] Yee Whye Teh. Dirichlet process. *Encyclopedia of Machine Learning*, 1063:280–287, 2010.

[65] Pierre C Delattre, Alvin M Liberman, and Franklin S Cooper. Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, 27(4):769–773, 1955.

[66] Abdellah Fourtassi and Emmanuel Dupoux. A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In *Conference on Computational Natural Language Learning (CoNLL)*, pages 191–200, 2014.

[67] Abdellah Fourtassi, Ewan Dunbar, and Emmanuel Dupoux. Self-consistency as an inductive bias in early language acquisition. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[68] Naomi Feldman, Thomas Griffiths, and James Morgan. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2208–2213, 2009.

[69] Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *NIPS*, pages 641–648, 2007.

[70] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *NIPS*, pages 1385–1392, 2005.

[71] Michael Heck, Sakriani Sakti, and Satoshi Nakamura. Iterative training of a DPGMM-HMM acoustic unit recognizer in a zero-resource scenario. In *SLT*, pages 57–63, 2016.

[72] Andrew Rosenberg and Julia Hirschberg. V-measure: a conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, pages 410–420, 2007.

[73] Michael Heck, Sakriani Sakti, and Satoshi Nakamura. Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero-resource scenario. *Procedia Computer Science*, 81:73–79, 2016.

[74] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[75] Steven Greenberg, Hannah Carvey, Leah Hitchcock, and Shuangyu Chang. Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31(3-4):465–485, 2003.

[76] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NIPS*, pages 6306–6315, 2017.

[77] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *NIPS*, pages 1878–1889, 2017.

[78] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. *arXiv preprint arXiv:1804.02812*, 2018.

[79] Siyuan Feng, Tan Lee, and Zhiyuan Peng. Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling. In *INTERSPEECH*, pages 1093–1097, 2019.

[80] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[81] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[82] Janet F Werker and Richard C Tees. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63, 1984.

[83] Charles Darwin. *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life (6th ed.)*. London: John Murray, 2009.

[84] Richard Wolfgang Semon. *The Mneme*. G. Allen & Unwin Limited, 1921.

[85] Bertrand Russell. *Analysis of mind*. London: George Allen and Unwin, 1921.

[86] Sheena A Josselyn and Susumu Tonegawa. Memory engrams: recalling the past and imagining the future. *Science*, 367(6473), 2020.

[87] Sheena A Josselyn, Stefan Köhler, and Paul W Frankland. Heroes of the engram. *Journal of Neuroscience*, 37(18):4647–4657, 2017.

[88] Leon G Reijmers, Brian L Perkins, Naoki Matsuo, and Mark Mayford. Localization of a stable neural correlate of associative memory. *Science*, 317(5842):1230–1233, 2007.

[89] Xu Liu, Steve Ramirez, Petti T Pang, Corey B Puryear, Arvind Govindarajan, Karl Deisseroth, and Susumu Tonegawa. Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature*, 484(7394):381–385, 2012.

[90] Jin-Hee Han, Steven A Kushner, Adelaide P Yiu, Hwa-Lin Liz Hsiang, Thorsten Buch, Ari Waisman, Bruno Bontempi, Rachael L Neve, Paul W Frankland, and Sheena A Josselyn. Selective erasure of a fear memory. *Science*, 323(5920):1492–1496, 2009.

[91] Steve Ramirez, Xu Liu, Pei-Ann Lin, Junghyup Suh, Michele Pignatelli, Roger L Redondo, Tomás J Ryan, and Susumu Tonegawa. Creating a false memory in the hippocampus. *Science*, 341(6144):387–391, 2013.

[92] Gisella Vetere, Lina M Tran, Sara Moberg, Patrick E Steadman, Leonardo Restivo, Filomene G Morrison, Kerry J Ressler, Sheena A Josselyn, and Paul W Frankland. Memory formation in the absence of experience. *Nature neuroscience*, 22(6):933–940, 2019.

[93] Stephen J Martin, Paul D Grimwood, and Richard GM Morris. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual Review of Neuroscience*, 23(1):649–711, 2000.

[94] Julie A Kauer and Robert C Malenka. Synaptic plasticity and addiction. *Nature reviews neuroscience*, 8(11):844–858, 2007.

[95] John Locke. *An essay concerning human understanding*. London: Thomas Basset, 1690.

[96] David Hume. *An enquiry concerning human understanding*. London: A. Millar, 1748.

[97] Arthur G Samuel. Lexical representations are malleable for about one second: evidence for the non-automaticity of perceptual recalibration. *Cognitive psychology*, 88:88–114, 2016.

[98] Peter D Eimas and John D Corbit. Selective adaptation of linguistic feature detectors. *Cognitive psychology*, 4(1):99–109, 1973.

[99] Dennis Norris, James M McQueen, and Anne Cutler. Perceptual learning in speech. *Cognitive psychology*, 47(2):204–238, 2003.

[100] Josh H McDermott. *Audition*. Oxford University Press, 2014.

[101] Johannes Maagaard Nielsen. *Agnosia, apraxia, aphasia: their value in cerebral localization*. New York, NY: Paul B. Hoeber, Inc., 1946.

[102] Sigmund Freud. *Zur auffassung der aphasien: eine kritische studie*. Leipzig: F. Deuticke, 1891.

[103] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

[104] Bin Wu, Sakriani Sakti, and Satoshi Nakamura. Incorporating discriminative DPGMM posteriorgrams for low-resource ASR. In *SLT*, pages 201–208, 2021.

[105] Sharon Goldwater, Mark Johnson, and Thomas L Griffiths. Interpolating between types and tokens by estimating power-law generators. In *NIPS*, pages 459–466, 2006.

[106] Jessica Maye, Janet F Werker, and LouAnn Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111, 2002.

[107] Bart De Boer and Patricia K Kuhl. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4(4):129–134, 2003.

[108] Bob McMurray, Richard N Aslin, and Joseph C Toscano. Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12(3):369–378, 2009.

[109] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[110] Roch Paulin Beapami, Ruth Chatfield, G Kouarata, and Andrea Waldschmidt. *Dictionnaire Mbochi-Français*. Point Noire, Congo: SIL-Congo, 2000.

[111] Luc Bouquiaux and Jacqueline MC Thomas. *Enquête et description des langues à tradition orale*, volume 1. Leuven, Belgium: Peeters Publishers, 1976.

[112] Jamison Cooper-Leavitt, Lori Lamel, Annie Rialland, Martine Adda-Decker, and Gilles Adda. Corpus base linguistic exploration via forced alignments with a light-weight ASR tool. In *Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poland, 2017.

[113] Ofir Press and Lior Wolf. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*, 2016.

[114] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[115] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[116] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *ICML*, pages 1310–1318, 2013.

[117] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*, 2017.

[118] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[119] Benjamin Van Niekerk, Leanne Nortje, and Herman Kamper. Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. In *INTERSPEECH*, pages 4836–4840, 2020.

[120] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs. RNN in speech applications. In *ASRU*, pages 449–456, 2019.

[121] Peter D Eimas, Einar R Siqueland, Peter Jusczyk, and James Vigorito. Speech perception in infants. *Science*, 171(3968):303–306, 1971.

[122] Josiane Bertoncini, Ranka Bijeljac-Babic, Sheila E Blumstein, and Jacques Mehler. Discrimination in neonates of very short CVs. *The Journal of the Acoustical Society of America*, 82(1):31–37, 1987.

[123] Sandra E Trehub. Infants' sensitivity to vowel and tonal contrasts. *Developmental Psychology*, 9(1):91–96, 1973.

[124] Philip J Swoboda, Philip A Morse, and Lewis A Leavitt. Continuous vowel discrimination in normal and at risk infants. *Child Development*, 47(2):459–465, 1976.

[125] Peter W Jusczyk. *Infant speech perception: A critical appraisal. Miller, J. L. and Eimas, P. D. (Eds.)*, pages 113–164. New York, NY: Psychology Press, 1982.

[126] Rebecca E Eilers and D Kimbrough Oller. The role of speech discrimination in developmental sound substitutions. *Journal of child language*, 3(3):319–329, 1976.

[127] Mary Skeel Abbs and Fred D Minifie. Effect of acoustic cues in fricatives on perceptual confusions in preschool children. *The Journal of the Acoustical Society of America*, 46(6B):1535–1542, 1969.

[128] Mary Anne Hanner. *Auditory discrimination and phonetic contexts in school age children.* PhD thesis, the Department of Speech Pathology, Eastern Illinois University, 1974.

[129] Peter W Jusczyk, Heather Copan, and Elizabeth Thompson. Perception by 2-month-old infants of glide contrasts in multisyllabic utterances. *Perception & Psychophysics*, 24(6):515–520, 1978.

[130] Joseph M Byrne, Cynthia L Miller, and Bonnie Hondas. Psychophysiologic and behavioral responsivity to temporal parameters of acoustic stimuli. *Infant Behavior and Development*, 17(3):245–254, 1994.

[131] Joanne L Miller and Peter D Eimas. Studies on the perception of place and manner of articulation: a comparison of the labial-alveolar and nasal-stop distinctions. *The Journal of the Acoustical Society of America*, 61(3):835–845, 1977.

List of publications

Journals

- Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura, "Tackling perception bias in unsupervised phoneme discovery using DPGMM-RNN hybrid model and functional load," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 348–362, 2021.

- Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura, "Modeling unsupervised empirical adaptation by DPGMM or DPGMM-RNN hybrid model to extract perceptual features for low-resource ASR," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 901-916, 2022.

International Conferences

- Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura, "Optimizing DPGMM clustering in zero resource setting based on functional load," in SLTU, 2018, pp. 1–5.

- Bin Wu, Sakriani Sakti, and Satoshi Nakamura, "Incorporating discriminative DPGMM posteriorgrams for low-resource ASR," in SLT, 2021, pp. 201–208.

Domestic Conferences

- Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura, "Using Functional Load for Optimizing DPGMM based Zero Resource Sub-word Unit Discovery,", Technical Report in IPSJ-SIG, Vol. 2018-SLP-125, No. 4, 1-2, Dec., 2018.
- Bin Wu, Sakriani Sakti, and Satoshi Nakamura, "MFCC-DPGMM Features for Enhancing Low-Resource ASR,", Technical Report in ASJ, Mar., 2021.