# Modeling Unsupervised Empirical Adaptation by DPGMM and DPGMM-RNN Hybrid Model to Extract Perceptual Features for Low-Resource ASR

Bin Wu, Sakriani Sakti, *Member, IEEE*, Jinsong Zhang, *Member, IEEE*, and Satoshi Nakamura, *Fellow, IEEE*

***Abstract*—Speech feature extraction is critical for ASR systems. Such successful features as MFCC and PLP use filterbank techniques to model log-scaled speech perception but fail to model the adaptation of human speech perception by hearing experiences. Infant perception that is adapted by hearing speech without text may cause permanent brain state modifications (engrams) that serve as a physical fundamental basis for lifetime speech perception formation. This realization motivates us to propose to model such an unsupervised adaptation process, where adaptation denotes perception that is affected or changed by the history of experiences, with the Dirichlet Process Gaussian Mixture Model (DPGMM) and the DPGMM-RNN hybrid model to extract perceptual features to improve ASR. Our proposed features extend MFCC features with posteriorgrams extracted from the DPGMM algorithm or the DPGMM-RNN hybrid model. Our analysis shows that the DPGMM and DPGMM-RNN model perplexities agree with infant auditory perplexity to support that the proposed features are perceptual. Our ASR results verify the effectiveness of the proposed unsupervised features in such tasks as LVCSR on WSJ and ASR on noisy low-resource telephone conversations, compared with the supervised bottleneck features from Kaldi in ASR performance.**

***Index Terms*—DPGMM, Zerospeech, unsupervised phoneme discovery, infant speech perception, low-resource ASR, engrams.**

## I. Introduction

SPEECH feature extraction can affect ASR performance. Such features as Mel-Frequency Cepstrum Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2] work well in ASR systems using mel-scaled and bark-spaced filterbanks [1], [2] 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.

### A. 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 [3]. 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. Experiences can leave "a permanent record ...written or engraved on the irritable substance" [4], and "past occurrences in the history of the organism as part of the causes of the present response" [5]. The term "permanent record" is coined as a "mnemic trace" or an "engram" by the evolutionary biologist Richard Semon [4], [5], who first introduced the concept to the scientific community.

Engram research of mnemic phenomena has recently become an exciting topic in neural science [6]. 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 permanent brain changes. Several generations failed for about ten decades until "engram renaissance" [7] started from the early 21th century, sparked by the development of molecular and circuit tools that probe and precisely manipulate

Manuscript received July 25, 2021; revised November 28, 2021 and January 9, 2022; accepted January 9, 2022. Date of publication February 10, 2022; date of current version March 3, 2022. This work was supported in part by JSPS KAKENHI under Grants JP21H05054 and JP21H03467. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tan Lee. *(Corresponding author: Sakriani Sakti.)*

Bin Wu is with the Nara Institute of Science and Technology, Ikoma 630-0192, Japan (e-mail: wu.bin.vq9@is.naist.jp).

Satoshi Nakamura is with the Nara Institute of Science and Technology, Ikoma 630-0192, Japan and also with the RIKEN Center for Advanced Intelligence Project (AIP), Ikoma 630-0192, Japan (e-mail: s-nakamura@is.naist.jp).

Sakriani Sakti is with the Japan Advanced Institute of Science and Technology, Nomi 923-1292, Japan, also with the Nara Institute of Science and Technology, Ikoma 630-0192, Japan, and also with the RIKEN Center for Advanced Intelligence Project (AIP), Ikoma 630-0192, Japan (e-mail: ssakti@jaist.ac.jp).

Jinsong Zhang is with the Beijing Language and Culture University, Beijing 100083, China (e-mail: jinsong.zhang@blcu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2022.3150220

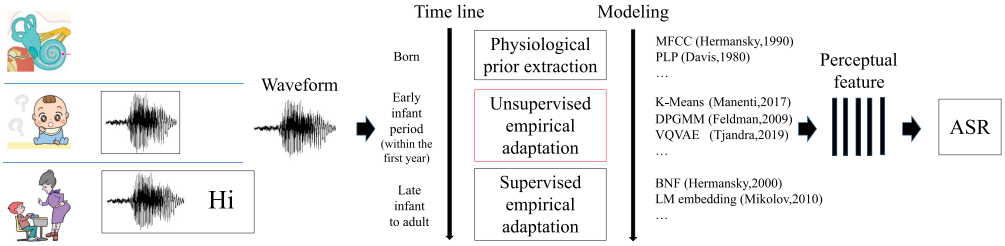



Fig. 1. Lifetime speech perception formation process. One initializes speech perception by auditory organs after birth, adapts it based on hearing speech without text in the early infant period, and adapts it again by learning experiences that connect speech with text from stages of late infant to adult. 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.

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 [8]. The tagged cells can be physically manipulated to make mice recall experiences without stimuli [9], disrupt brain records as if such experiences never happened [10], or even implant "fake" memories of non-existing experiences [11], [12]. The endurance of engram changes was verified by measuring the strength of engram cell connections [13], [14].

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 [15].

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" [5]. Russell borrowed the concepts of mnemic phenomena and engrams from Semon. He elaborated the essence of perception in the tradition of Locke [16] and Hume [17], 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 [6].

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 [18]; such adaptations include selective adaptation [19] that lasts for minutes, lexicon adaptation [20] for hours or days (after stimuli of minutes), and language learning adaptation for months or years [21].

### *B. Infant Learning Experiences to Establish Lifetime Perception*

Speech perception is adapted through language learning experiences (Fig. 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 [22].

The primary stage of language learning starts when an infant's "psychological urge" [21] 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" [21] 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" [23]—may cause perception impairment [24], 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 [21].

After the primary stage, an infant enters the second stage of language learning called the vocabulary spurt [25] 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 [21], 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 [25].

Modeling the speech perception formation process (Fig. 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 [1] and PLP [2] features; to model the supervised empirical adaptation that learns from speech and text, we can extract supervised features such as bottleneck features (BNF) [26] or language embeddings [27]. However, modeling unsupervised empirical adaptation in the infant period has been less explored for ASR applications (highlighted in red rectangle in Fig. 1), especially for Large Vocabulary Continuous Speech Recognition (LVCSR) or low-resource ASR.

### *C. 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) [28] 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 [29]–[31]. 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 [32], neural network features from autoencoder [33], ABnet [34], and VQ-VAE [35], parametric clustering features from GMM [35] and K-means [35], [36], and nonparametric clustering features from DPGMM trained with Gibbs sampling [29], [30] and variational inference [37], [38]. DPGMM also worked in spoken term detection [39], but it was rarely applied in ASR, especially in LVCSR [40] or low-resource ASR that we will tackle in this paper.

The DPGMM is interpretable as a graphical model [28] 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 [41] that contain more sophisticated explainable linguistic factors, including lexicon or grammar priors [42].

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 [42]. 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 [43].

This stream of literature aims to use model simulation to illustrate infant distributional learning [44], [45] during phoneme category acquisition and to provide evidence for mechanisms [42], [44] to explain the developmental changes [3], [42], [46] 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 [44] to demonstrate an infant's sensitivity to the statistical distribution of speech sounds. Boerl and Kuhl *et al.* used GMM with the EM algorithm [45] 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 [46] 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 [42]. Feldman's finding of the interactive learning process of the infants using DPGMM is well referenced by cognitive science, psychology, and infant language acquisition.

### *D. Modeling Unsupervised Empirical Adaptation by DPGMM-RNN Hybrid Models for ASR*

However, DPGMM fails to model the temporal order of speech features [47], 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 [48]. 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 and creates small, random cluster segments inside a phoneme. This is DPGMM's "fragmentation problem" [49].

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 [49], [50]. DPGMM tends to generate more clusters than the number of phonemes in any human language [30], [50] when it struggles to discriminate between complex phonemes with higher resolution.

We propose to use the DPGMM-RNN hybrid model [49], 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

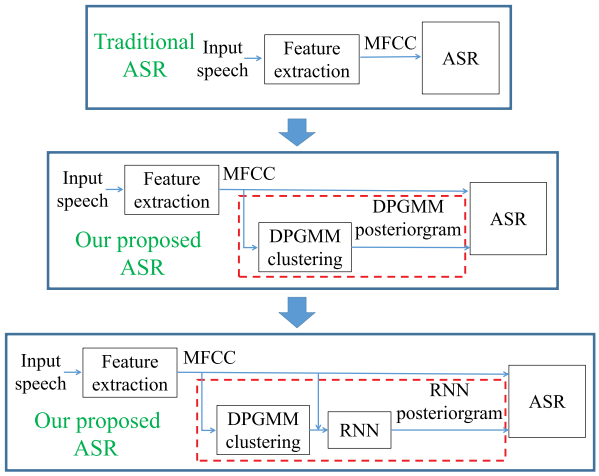



Fig. 2. Proposed feature extension by concatenating an MFCC feature with a DPGMM posteriorgram (from the DPGMM clustering algorithm) or with an RNN posteriorgram (from the DPGMM-RNN hybrid model) in feature extraction for ASR. A posteriorgram is a vector whose $k$-th dimension represents the probability that an observed frame belongs to the $k$-th cluster.

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 [51] and the v-measure [52]. It also reduced the number of clusters of DPGMM [49] and overperformed DPGMM in unsupervised phoneme discovery on datasets from Zerospeech 2019 with an ABX discrimination test at a moderate bitrate [49].

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 [5] 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 auditory 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. 1), where these perceptual features extend MFCC features with DPGMM or DPGMM-RNN posteriorgrams by concatenation (Fig. 2).

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 [53] and WSJ [54] (a widely used dataset for LVCSR) and on the low-resource corpora of Mboshi [55] and Javanese [56] (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 [57].
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 further compare the phoneme categorization of the models with that of infants. Our analysis is based on our idea that for a model representation that categorizes phonemes well, the entropy of the representation of a phoneme should be small and the relative entropy of the representations of different phonemes should be large.

**Algorithm 1:** Gibbs sampling for DPGMM (Fig. 3) given hyperparameters $\alpha$ and $\beta$ and observed data $x$.

Randomly initialize cluster indicator $z = z_1, \ldots, z_n$
**for** Iteration $iter = 1, 2, \ldots$ **do**
  Sample $\pi' \sim p(\pi|z, \alpha)$ by (1),
  $\pi_1, \ldots, \pi_K, \pi^*_{K+1}|z, \alpha \sim \mathrm{Dir}(n_1, n_2, \ldots, n_K, \alpha)$
  Sample $\mu,' \Sigma' \sim p(\mu, \Sigma|z, \beta, x)$ by Eq. (2),
  $\mu_k, \Sigma_k|z, \beta, x \sim \mathrm{NIW}(\mu_0^{(k)}, \lambda^{(k)}, \Sigma_0^{(k)}, \nu^{(k)})$
  Sample $z'_i \sim p(z_i|\pi,' \mu,' \Sigma,' x_i)$ by Eq. (4),
  $z_i|\pi, \mu, \Sigma, x_i \sim \pi_k p(x_i|\mu_k, \Sigma_k)/p(x_i)$
  Update $z = (z_1,' \ldots, z'_n)$.
**end for**

## II. Method

### A. Feature Extensions by Concatenating MFCC Features With DPGMM or RNN Posteriorgrams for ASR

We propose to extend the MFCC features with DPGMM posteriorgrams or DPGMM-RNN posteriorgrams by concatenation (Fig. 2) to improve ASR, where

1) the MFCC features represent acoustic features,
2) the DPGMM generates DPGMM posteriorgrams after adaptation of DPGMM parameters with MFCC features,
3) 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,
4) 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). In the following subsections, we introduce the DPGMM and DPGMM-RNN hybrid model for generating posteriorgrams.

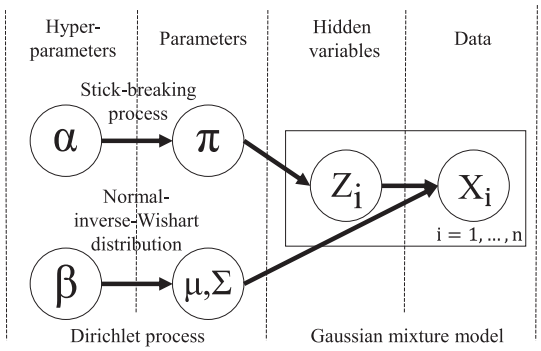



Fig. 3. Graphical model of Dirichlet Process Gaussian Mixture Model (DPGMM) generates parameters of weights ($\pi = \pi_1, \ldots, \pi_k, \ldots$), means, and variances ($(\mu, \Sigma) = (\mu_1, \Sigma_1), \ldots, (\mu_k, \Sigma_k), \ldots$) of Gaussians from stick-breaking process (with concentration parameter $\alpha$) 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 $\mu_k$ and variance $\Sigma_k$ indicated by hidden variable $Z_i = k$. Rectangular box, with $(Z_i, X_i)$ inside, is a simplified notation of all $n$ data points (features) with their indicator hidden variables $((Z_1, X_1), \ldots, (Z_i, X_i), \ldots, (Z_n, X_n))$.

### B. DPGMM Clustering Algorithm

*1) DPGMM's Definition as a Graphical Model:* 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. 3) with the following generative process.

- It generates mixture weights $\{\pi_k\}_{k=1}^{\infty}$ from a stick-breaking process [58] with concentration parameter $\alpha$;
- it generates means and variances $\{\mu_k, \Sigma_k\}_{k=1}^{\infty}$ from normal-inverse-Wishart (NIW) distribution with a belief of mean $\mu_0$, a belief of variance $\Sigma_0$, a belief-strength of mean $\lambda$, and a belief-strength of variance $\nu$; 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 $\mu_k$ and variance $\Sigma_k$ indicated by the hidden variable $Z_i = k$.

We summarize this generative procedure for the graphical model of $\mathrm{DPGMM}(\alpha, \mathrm{NIW}(\mu_0, \lambda, \Sigma_0, \nu))$ and describe the dependency relations among the random variables of the model in Fig. 3.

*2) DPGMM Training by Gibbs Sampling:* Given model $\mathrm{DPGMM}(\alpha, \mathrm{NIW}(\mu_0, \lambda, \Sigma_0, \nu))$, data $\{x_i\}_{i=1}^{n}$, and any indicator initialization, we get posteriorgram $p(z_i|x_i)$ by iteratively inferring from the Gibbs sampling (Algorithm 1) with the following steps until convergence.

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

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

where K is the number of 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, counted by hidden indicator variables $z = z_1, \ldots, z_n$.

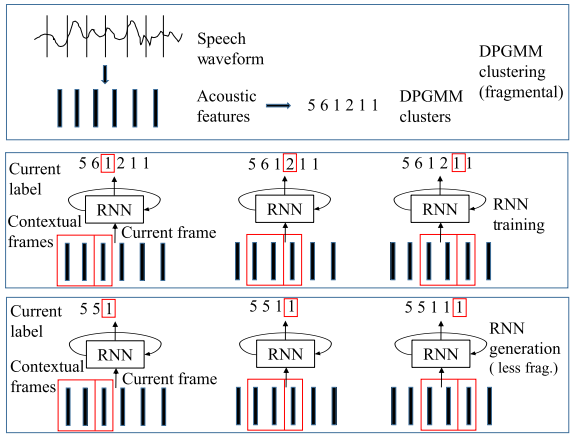



Fig. 4. Three steps for constructing DPGMM-RNN hybrid model. The RNN target can be a DPGMM cluster label for unsupervised phoneme discovery or a DPGMM posterior vector for unsupervised feature extraction.

Second, we update the mean and variance for each Gaussian cluster $k$ by sampling a normal-inverse-Wishart distribution [59] after observing data $x$:

$$\mu_k, \Sigma_k|z, \beta, x \sim \mathrm{NIW}(\mu_0^{(k)}, \lambda^{(k)}, \Sigma_0^{(k)}, \nu^{(k)}), \quad (2)$$

where $\mu_0^{(k)}$, $\lambda^{(k)}$, $\Sigma_0^{(k)}$, and $\nu^{(k)}$ are the updated parameters for the $k$-th cluster after seeing the data [59]:

$$\begin{aligned}
\mu_0^{(k)} &= \frac{\lambda}{\lambda + n_k} \cdot \mu_0 + \frac{n_k}{\lambda + n_k} \cdot \bar{x}_k \\
\lambda^{(k)} &= \lambda + n_k \\
\nu^{(k)} &= \nu + n_k \\
\Sigma_0^{(k)} &= \Sigma_0 + S_k + \frac{\lambda n_k}{\lambda + n_k} (\bar{x}_k - \mu_0)(\bar{x}_k - \mu_0)^T \\
&\text{with} \\
\bar{x}_k &= \frac{1}{n_k} \sum_{i=1, z_i=k}^{n} x_i; S_k = \sum_{i=1, z_i=k}^{n} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T.
\end{aligned} \quad (3)$$

Third, we update the hidden variables by sampling the posterior distribution:

$$p(z_i = k|\pi, \mu, \Sigma, x_i) = \frac{\pi_k p(x_i|\mu_k, \Sigma_k)}{p(x_i)} \propto \pi_k p(x_i|\mu_k, \Sigma_k). \quad (4)$$

### C. DPGMM-RNN Hybrid Model

The DPGMM-RNN hybrid model uses RNN to refine the DPGMM clusters or posteriorgrams in the following three steps (Fig. 4):

- DPGMM clustering: we apply the DPGMM clustering algorithm to get a cluster label (or a posterior vector) for each feature frame.
- 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 reconstruction: we use RNN to get the posterior vector framewisely by inputting the speech segment. The dimension of the maximum probability in the posterior vector is chosen as the output cluster label.

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.

### D. Conditional Perplexity and KL Divergence

We analyze the fragmental level of the generated representations from the DPGMM or DPGMM-RNN hybrid model with the conditional perplexity of cluster given phoneme [49] that reflects the average number of cluster types corresponding to one phoneme type. We define the conditional perplexity by the exponential of the conditional entropy [51] of the cluster representation ($C$) given the phoneme truth ($T$) with base 2.

$$ppl(C|T) = 2^{H(C|T)}, \tag{5}$$

$$\begin{aligned} H(C|T) &= \sum_t p(t) H(C|T=t) \\ &= -\sum_t p(t) \sum_c p(c|t) \cdot log p(c|t) \\ &= -\sum_t \frac{n_t}{n} \sum_c \frac{n_{ct}}{n_t} \cdot log \frac{n_{ct}}{n_t}, \end{aligned} \tag{6}$$

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.

We analyze the discriminability ($D$) between the representations of a phoneme pair ($t_1$ and $t_2$) by the KL divergence between the conditional distributions of cluster representation ($C$) given the phoneme. We compute the conditional distributions using the relative frequencies of the clusters.

$$\begin{aligned} D(t1, t2) &= KL(P(C=c|T=t_1)||P(C=c|T=t_2)) \\ &= \sum_c p(c|t_1) log \frac{p(c|t_1)}{p(c|t_2)}, \end{aligned} \tag{7}$$

## III. Dataset and Experiment Setup

### A. Datasets and Their Divisions

*1) TIMIT:* We analyzed the models on the TIMIT corpus [53] of English read speech because it includes reliable and detailed phoneme annotations. We followed the official division [53] of a training set of 3.14 hours, a development set, and a complete test set of 1344 utterances.

*2) WSJ:* We checked the LVCSR performance on the WSJ corpus [54] of the English speech. We followed the official division [54] 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.

*3) Mboshi:* We further experimented on a low-resource African read corpus of Mboshi [55] 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 [60] and a fieldwork-oriented Bouquiaux and Thomas's corpus [61].

TABLE I
Statistics of Low-Resource Mboshi Read Speech Datasets [55] of Three Speakers

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

The Mboshi dataset [56] 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 I) 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 III-B2). Table I summarizes the statistics of the Mboshi dataset.

*4) Javanese:* We attempted some challenging experiments on a low-resource Indonesia telephone conversational corpus of Javanese [56] 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 [56] that 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 III-B2). Table II summarizes the statistics of the Javanese dataset.

### B. Feature Extraction

*1) Acoustic Feature Extraction:* We followed Kaldi [57] using a 39-dimensional MFCC+$\Delta$+$\Delta\Delta$ (25-ms frame size and 10-ms frame shift) with mean and variance normalization (CMVN)

TABLE II
STATISTICS OF LOW-RESOURCE JAVANESE TELEPHONE DATASETS [56]. 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 OR DEVELOPMENT SET

| Javanese | #Hours | #Utterances | #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 |

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.

*2) 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 [62] 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.

*3) 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 [30], [49], [63]. 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 [49] and fed the RNN with the MFCC feature chunk of a center frame binding with eight left and eight right adjacent frames, the optimal size of context according to earlier work [49]. The RNN mapped the feature chunk to the posteriorgram of the center frame. We used an RNN of a 5-layer BiLSTMs [49] 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. We implemented RNN using Pytorch. We trained RNN to minimize the loss function of the mean square error (MSE) between the softmax layer outputs and the DPGMM posteriorgrams using 20 epochs with a learning rate of 0.001 and a batch size of 256.

*4) BNF Feature Extraction:* We compared our proposed unsupervised DPGMM features with supervised BNF features from Kaldi. The BNF feature extraction requires reliable alignments of the starting and ending times of each phoneme to work well. We obtained the ASR alignments and extracted the BNF features of WSJ and TIMIT with the official Kaldi scripts (run.sh and run_bnf.sh) without modification except for changing the 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 features of Mboshi and Javanese following the Kaldi scripts of TIMIT because these datasets have similar amounts of data.

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) [57].

*5) VQCPC Feature Extraction:* We compared our proposed unsupervised DPGMM features with the unsupervised VQCPC feature. The VQCPC model attained the top result for ABX error rate at Zerospeech 2020 [66]. The VQCPC model had a slightly better ABX error rate with a lower bitrate compared with the DPGMM [31] on the same Zerospeech English test dataset.

We used the open-source code of VQCPC from an identical model structure [66] to extract the VQCPC feature from the same MFCC feature with which we extracted DPGMM-based features. The only modification of the default parameters is to close the subsampling before the encode layer by setting the stride of CNN from 2 to 1 and the kernel size of CNN from 4 to 3 so that we could obtain the same number of frames of the output VQCPC features as the input MFCC features. We extracted the VQCPC feature from the model of the last training epoch of the default setup.

The VQCPC model encodes the input feature with a convolutional layer and four linear layers using ReLU activations and layer normalizations, followed by a linear bottleneck layer of 64 dimensions (to extract the feature) and a vector quantization layer of 512 codes. Finally, an LSTM layer summarizes the encoded discrete representation into a context vector, with which the model is trained to discriminate future codes from negative examples drawn from other utterances.

### C. ASR System

We used Pytorch to implement an attentional encoder-decoder ASR system [64] from scratch that 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 detailed structure is summarized in Table III.

The ASR system maps sequences of speech features to sequences of characters. We processed the provided transcriptions into the characters following an earlier work [64] with a character set that includes a, b, c,..., z, <space >, <comma>, <period >, <apostrophe>, <unk>, <sos>, and <eos> (where < sos> and <eos> denote start-of-sentence and end-of-sentence tokens) for

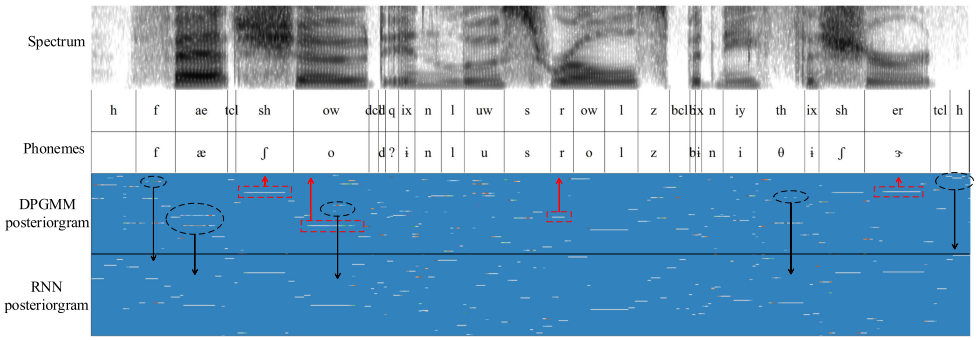



Fig. 5. Phoneme recognition (red rectangles) and fragmentation problem (black circles) of posteriorgrams. Utterance "Fat showed in loose rolls beneath the shirt" with id FADG0_SI1909 from TIMIT test set shows posteriorgrams from DPGMM clustering algorithm (DPGMM posteriorgram) and DPGMM-RNN hybrid model [49] (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.

TABLE III
ARCHITECTURE OF ATTENTIONAL ENCODER-DECODER ASR SYSTEM. A → B DENOTES NEXT LAYER OF LAYER A IS LAYER B. PBILSTM DENOTES A PYRAMID BIDIRECTIONAL LSTM [64]; 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 [65], 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 AT EACH LAYER

| **Module** | **Cascaded module layers** |
|---|---|
| Encoder | FC-512 → ReLU → Dropout → 3-layer pBiLSTM-256 (reduced half of frames per layer) |
| Decoder pre-net | EMBED-256 → Dropout |
| Decoder [65] | (Pre-net output + Prev. decoder output) Single-layer LSTM-512 → Dropout → Contextual FC-256 → Tanh |
| Decoder post-net | Softmax |
| MLP attention | FC-256 → Tanh |

WSJ, TIMIT, and Javanese, and with a character set that includes the additional UTF-8 characters of á, é, í, ó, ú, $\varepsilon$, $\acute{\varepsilon}$, $\omega$, and $\acute{\omega}$ for Mboshi.

The setups of the ASR include a dropout probability of 0.05, a label smoothing ratio of 0.05, a learning rate of 0.001 (which halved whenever the training loss successively increases for more than three epochs), and a beam size of 10.

All ASR results reported in this paper are from this ASR system without any pronunciation dictionaries or language models in the decoding process. We evaluated the ASR results on the test sets of four datasets following the splits of training, development, and test sets described in the dataset section.

We used the transformer-based ASR from ESPnet with a language model on WSJ SI-284 for LVCSR. We used the identical python environment and the transformer structure following the configuration files of "train_pytorch_transformer.yaml" and "decode_pytorch_transformer.yaml" in ESPnet. The transformer-based ASR, described by the "RESULTS.md" file, used the joint CTC training with a weight of 0.3 to CTC loss. The decoding process used a language model trained from about 65,000 words and a beam size of 10. The input layer of the transformer's encoder used a 2D convolutional layer. The transformer structure includes 12 hidden encoder layers with 2048 units in each layer, 6 hidden decoder layers with 2048 units in each layer, and 4-headed attention with 256 units.

## IV. RESULT

### A. Discriminative Posteriorgram and Fragmentation Problem

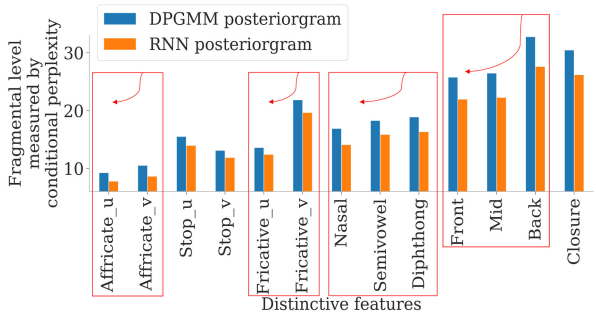
We concatenated the MFCC features with the posteriorgrams from the DPGMM clustering algorithm or the DPGMM-RNN hybrid model. We describe the characteristics of these posteriorgrams using an utterance from the TIMIT test set.

Fig. 5 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.

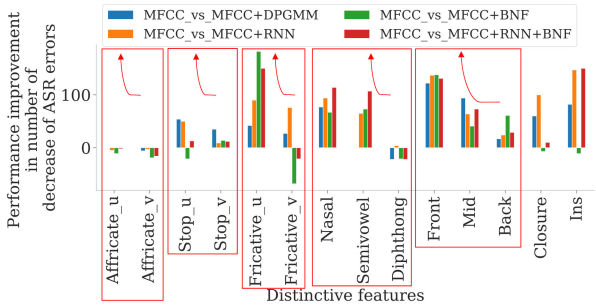
Fig. 5 also shows that the RNN posteriorgram can relieve the fragmental problems from the DPGMM posteriorgram on such phonemes with complex acoustics as fricatives that contain noise-like high-frequency components (see the black circles) and that the RNN posteriorgram discovered more phonemes.

### B. 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 [57] 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. 6) is the difference of




(a) Fragmental level measured by conditional perplexity




(b) ASR improvement measured by error decrease

Fig. 6. Fragmental levels and ASR improvements of distinctive features on TIMIT test set. Upper subfigure (a) conditional perplexity of cluster given phonemes [49] that shows fragmental level of posteriorgrams from DGPMM algorithm (DPGMM posteriorgram) and DPGMM-RNN hybrid model [49] (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 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.

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 [49] 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. 6 shows the following results.

- Unvoiced consonants, less fragmental than voiced ones, tend to have more ASR improvement.
- From Fig. 6, we observe an exception where an unvoiced stop, "stop_u," is more fragmental than a voiced stop, "stop_v"; here, the unvoiced stop has better ASR performance than the voiced stop. An unvoiced stop usually has a longer and more obvious aspiration subsegment compared with a voiced stop, which makes the unvoiced stop more segmental with higher perplexity. Such an aspiration subsegment can serve as a landmark [67] of an unvoiced stop to make it easier for the ASR to identify.
- 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.
- The RNN posteriorgram relieves the fragmental problem of the DPGMM posteriorgram [49] (Fig. 5), indicated by decrease of fragmental level measured by conditional perplexity for each distinctive feature (Fig. 6(a)). 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. 6(b)).
- 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. 6(b) and 5).
- 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.

### C. *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).

In our experiments, we obtained 99 DPGMM clusters with 99-dimensional DPGMM posteriorgrams. 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 [70]–[72], and the DPGMM-RNN hybrid model outperformed the DPGMM clustering algorithm at discriminating and identifying phonemes [49]. Table V shows that the RNN posteriorgram (RNN) worked better than the DPGMM posteriorgram (DPGMM) in ASR, but neither reached the ASR performance of MFCC.

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. 5) that can be compensated by MFCC features. Table V 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 [40].

Table V shows that the feature extension with the RNN posteriorgram (MFCC+RNN) achieved a lower CER than that

TABLE IV
MAPPING FROM DISTINCTIVE FEATURE (ARTICULATORY FEATURE) TO PHONEMES. WE REPRESENT PHONEMES AS 39 TIMIT PHONEMIC SYMBOLS USED BY THE KALDI RECIPE. WE ALSO INCLUDED INTERNATIONAL PHONETIC ALPHABET (IPA) CHARACTERS OF TIMIT PHONEME SYMBOLS. HERE, 'STOP_U' DENOTES AN UNVOICED STOP, 'STOP_V' DENOTES A VOICED STOP, AND 'ʃ(ʒ)' MEANS ʃ AND ʒ ARE REPRESENTED AS THE IDENTICAL TIMIT PHONEMIC SYMBOL ('SH')

| Feature | Phoneme | IPA |
|---|---|---|
| Stop_u | p t k | p t k |
| Stop_v | b d g | b d g |
| 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) |

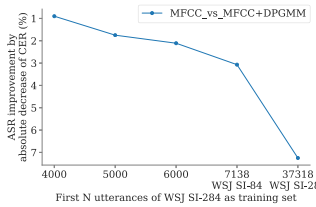
TABLE V
LVCSR PERFORMANCE ON WSJ. WE COMPARED MFCC FEATURES, DPGMM POSTERIORGRAMS, RNN POSTERIORGRAMS, AND THEIR CONCATENATIONS ON OUR ATTENTIONAL ENCODER-DECODER ASR SYSTEM, ALONG WITH TWO BASELINES [68], [69], BY CHARACTER ERROR RATES (CERS) WITHOUT PRONUNCIATION DICTIONARIES OR LANGUAGE MODELS IN DECODING PROCESS. WE TRAINED THE ASR MODEL ON THE WSJ SPEECH CORPUS [54] WITH THE TRAINING SET OF WSJ SI-84 (ABOUT 15 HOURS) OR WSJ SI-284 (ABOUT 80 HOURS), CHOSE THE BEST MODEL ON THE DEVELOPMENT SET OF DEV93, AND EVALUATED THE BEST MODEL ON THE TEST SET OF EVAL92 TO GET THE CER RESULTS. 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% [68]

| **System with feature** | **WSJ SI-84 CER(%)** | **WSJ SI-284 CER(%)** |
|---|---|---|
| Baseline ASR1 MEL [68] | 17.01 | 8.17 |
| Baseline ASR2 MEL [69] | 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 |

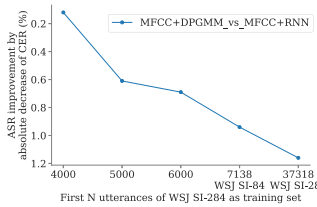
with the DPGMM posteriorgram (MFCC+DPGMM); both of the proposed feature extensions outperformed the MFCC feature (MFCC). The WERs were consistent with CERs; our proposed feature extension (MFCC+RNN) achieved a WER of 15.3%, compared with 18.2% in a previous work [68], on the WSJ SI-284 set with an encoder-decoder ASR without a language model.

We used the transformer-based ASR from ESPnet with a language model (as described in Section III-C of ASR system) for our experiments. We obtained CER (WER) of 3.1% (6.4%), 2.7% (5.7%), and 2.4% (5.4%) for the MFCC feature (MFCC) and the proposed features (MFCC+DPGMM and MFCC+RNN) on WSJ SI-284.

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




(a) ASR improvement from MFCC to MFCC+DPGMM




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

Fig. 7. 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 built ASR models with training sets of the first N utterances of the WSJ SI-284 set, an identical development set of dev93, and an identical test set of eval92, 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%.

on the WSJ SI-84 set (0.9% and 1.75% respectively in Table V). 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. 7).

Fig. 7(a) shows that extending the MFCC feature with the DPGMM posteriorgram (MFCC_vs_MFCC+DPGMM) improved the ASR performance more with less data. Fig. 7(b) shows that enhancing the DPGMM posteriorgram with the RNN posteriorgram (MFCC+DPGMM_vs_MFCC+RNN) improved the ASR performance more with less data.

When we examined whether proposed features are promising for low-resource ASR with small datasets, we found that absolute CER improvements are 0.9% and 1.75% and relative CER improvements are 13.7% and 10.54% for the datasets of WSJ SI-284 and WSJ SI-84 from the proposed features (MFCC+DPGMM) to the MFCC features (MFCC). The absolute and relative ASR improvements of (1.75%, 2.11%, 3.07%, 7.25%) and (10.54%, 11.26%, 11.29%, 16.38%) are consistent in their tendency to increase when trained with small datasets of less than 15 hours, including WSJ SI-84 as well as the first 6000, 5000, and 4000 utterances of WSJ SI-284.

### D. 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.,

TABLE VI
ASR PERFORMANCE ON LOW-RESOURCE CORPORA. 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 |

Javanese). We verified the effectiveness of our proposed features on the low-resource ASR.

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 VI shows that the ASR on Mboshi outperformed TIMIT.

The Javanese telephone conversation dataset is challenging for low-resource ASR. 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 VI shows lower ASR performance on Javanese than TIMIT or Mboshi.

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

Table VI 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 our DPGMM-based features with the unsupervised features of the VQCPC feature and a K-Means one-hot feature on the Javanese low-resource ASR. The VQCPC-extended feature (MFCC+VQCPC) attained a CER (WER) of 50.12% (64.07%), compared to 51.68% (64.75%) by the DPGMM-extended feature (MFCC+DGPMM) and 48.19% (61.37%) by the RNN-extended feature (DPGMM+RNN). The DPGMM model was found to suffer from the fragmental problem due to the noisy spontaneous Javanese speech. The VQCPC model without a Gaussian model constraint has a better generalization power than DPGMM, meanwhile it brings the risk of learning the noise into the feature. The K-means one-hot feature (MFCC+K-Means) gained a CER (WER) of 53.07% (68.30%). Our proposed RNN feature performed better than the VQCPC and K-means features for low-resource ASR.

The RNN of our DPGMM-RNN hybrid model used five hidden layers with a size of 512, resulting in a high computational cost that makes RNN difficult to use. Therefore, we reduced the number of parameters of RNN by decreasing the number of hidden layers, using 5, 4, and 3, and decreasing the size of the hidden layers using 512 and 256. We found that the CER (WER) of the RNN feature (MFCC+RNN) ranged from 48.19% (61.37%) to 50.13% (64.41%) compared to the DPGMM feature (MFCC+DPGMM, 51.68% (64.75%)).

The results reveal that reducing the number of RNN parameters slightly degrades the ASR performance of the RNN feature. Nevertheless, with a smaller parameter size (a 3-layer RNN; 256 hidden states in each layer), we can still achieve a better CER (WER), at 49.36% (62.64%), than that of the DPGMM feature.

A comparison of the RNN feature (MFCC+RNN) and DPGMM feature (MFCC+DPGMM) using the same ASR would not be fair because the extra parameters of RNN from the DPGMM-RNN hybrid model can contribute to ASR improvement. We increased the parameters of the ASR encoder of the DPGMM feature for a fairer comparison. We increased the number of layers to 3, 4, and 5 and increased the size of layers to 256 and 512. We found the performance of the DPGMM feature (MFCC+DPGMM) achieved a higher CER (WER), ranging from 48.76% (61.96%) to 50.13% (64.41%) compared to the DPGMM feature (MFCC+DPGMM) at 51.68% (64.75%) and the RNN feature (MFCC+RNN) at 48.19% (61.37%) when the encoder becomes larger.

Increasing the number of parameters of the ASR encoder makes the DPGMM feature perform better. But the performance of the RNN feature is better than the best performance of the DPGMM feature with a parameter-increased ASR. We consider that a whole system includes the feature-extraction model and the ASR system. Following the ASR parameter increment strategy for the DPGMM feature, when the parameter numbers of the whole systems for the RNN feature and the DPGMM feature are similar, the RNN feature also has a better ASR performance than the DPGMM feature.

### *E. Comparison and Combination With Supervised BNF in ASR*

Both supervised BNF [26] and unsupervised DPGMM-RNN features [49] 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 to work well; Kaldi [57] is state-of-art for this purpose.

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 VII 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 VII shows that the combination of MFCC, RNN, and BNF features (MFCC+RNN+BNF) worked best on

TABLE VII
ASR PERFORMANCE OF UNSUPERVISED AND SUPERVISED FEATURES. WE COMPARED UNSUPERVISED FEATURE EXTENSION WITH RNN POSTERIORGRAMS [49] (MFCC+RNN) WITH SUPERVISED FEATURE EXTENSION WITH BNF FEATURES [26] (MFCC+BNF). FOR WSJ AND TIMIT, WE USED KALDI'S [57] 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** | **Mboshi** | **TIMIT** | **WSJ SI-284** |
|---|---|---|---|---|
| Data amount (hours) | 2.88 | 2.00 | 3.14 | 81.25 |
| Data style | Spontaneous | Read | Read | Read |
| Recording devices | TEL/MOB | MIC | MIC | MIC |
| | **CER/WER (%)** | **PER (%)** | **PER (%)** | **CER/WER (%)** |
| MFCC | 53.23 / 66.44 | 22.67 | 23.92 | 6.57 / 16.96 |
| MFCC+RNN | 48.19 / 61.37 | **20.67** | **22.38** | 5.55 / 15.26 |
| MFCC+BNF [57] | 47.47 / 60.91 | 21.93 | 23.04 | 5.00 / 14.21 |
| MFCC+RNN+BNF | **43.23 / 57.31** | 21.26 | 22.53 | **4.48 / 12.74** |

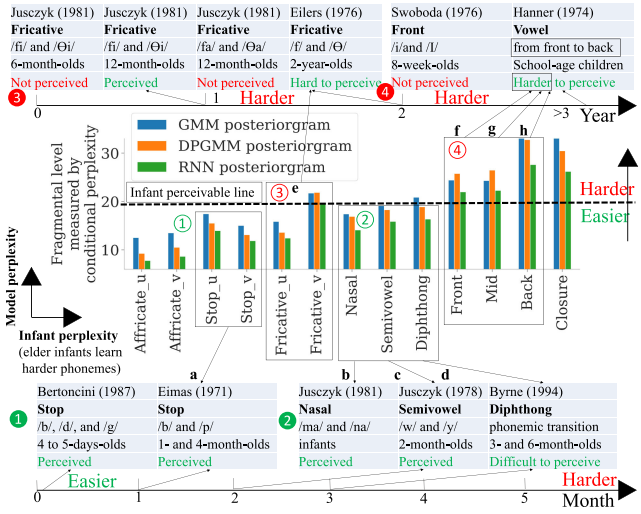



Fig. 8. Relation between DPGMM model perplexity on TIMIT corpus and infant perceptual perplexity by auditory experiments. Circled numbers denote degrees of perplexity, including DPGMM and DPGMM-RNN model perplexity vertically and infant perceptual perplexity horizontally. An infant-perceivable line is arbitrarily set to divide distinctive features that are easy (green) or hard (red) for infants to discriminate within the first year. We include a result of the GMM algorithm (GMM) from sklearn by setting the number of clusters to be identical to that of the DPGMM for comparison with the proposed DPGMM-based features (DPGMM and RNN).

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 VII 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. Several factors make the performances of features extracted from the BNF and RNN models different. First, both BNF and RNN models have five hidden layers, where the hidden size of RNN is 512 and that of BNF is 1024. BNF and RNN have a similar number of parameters. Second, the target for RNN training is 99 types of DPGMM clusters, while that for BNF training is 1896 types of HMM states. For small datasets of only a few hours, the training set may have insufficient samples to capture the complex distribution of a test set with many target types. Third, an RNN that uses a recurrent structure can better capture the temporal information than the feed-forward neural network used in Kaldi's implementation of 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 compared the results of our RNN-based ASR with the results of the transformer-based ASR without LM on WSJ SI-284 from the "RESULTS.md" file of ESPnet. We found the following results: 1) The MEL feature (with CER (WER) of 5.57% (15.49%)) performed better than the MFCC feature on our current baseline RNN-based ASR. The proposed feature (MFCC+RNN) is better than the MEL feature. 2) Our current RNN-based ASR (with fewer parameters) performed better than the transformer-based ASR of the base model ("Base Model" with CER (WER) of 5.7% (15.5%)) from ESPnet using the MEL feature. 3) Our current RNN-based ASR using the MFCC+RNN+BNF feature could perform better than a transformer-based ASR of the large model ("Big Model" with CER (WER) of 5.3% (14.0%))) using the MEL feature with the current state-of-the-art (SOTA) transformer structure of ESPnet (without the method of averaging of the ASR models of the last ten epochs).

## V. DISCUSSION

### A. 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 [73] and different from adult perception [3]. 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. 8) shed light on this question.

We define DPGMM perplexity of phonemes as conditional perplexity of DPGMM clusters given the phonemes [49]; 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. 8) 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/) [74]. 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/ [75]. Stops are among the easiest and the earliest distinctive features perceived by infants.

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 [76]. 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 [77].

The DPGMM perplexity of fricatives is high among the consonants. Fricatives /f/ and /$\theta$/ are fragmental with high perplexity and are frequently observed in individual utterance examples (Fig. 5) of DPGMM clustering [49]. Eimas *et al.* found that 6- and 12-month-olds cannot discriminate /fa/ and /$\theta$a/; only 12-month-olds can discriminate /fi/ and /$\theta$i/ [78]. The contrast of /f/ and /$\theta$/ is difficult for toddlers as well. Eilers and Oiler reported on 2-year-olds [79]; Abbs and Minifie reported on preschool children from 3- to 5-year-olds [80].

The DPGMM perplexity becomes higher from front vowels to back vowels. Swoboda *et al.* showed that 8-week-olds cannot categorize front vowels [77]. The accurate discrimination of vowels by school-age children, in the phonemic environment of /r/, is ranked roughly from front to back [81].

The DPGMM-RNN perplexity is smaller than the DPGMM perplexity in semivowels, diphthongs, and nasals, because DPGMM does not involve temporal order modeling [48], 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 [82]. Byrne *et al.* found that 3- and 6-month-olds can discriminate interphonemeic transitions inside a diphthong [83]. Nasals (/ma/ and /na/) can be distinguished by formant transitions [78], [84].

Our further analysis (Fig. 6(b)) suggests a potential causal relation between DPGMM computational perplexity and DPGMM ASR performance. We found some positive association between lower perplexity and higher ASR performance of the distinctive features. We also observed a contradictory tendency between "stop_u" and "stop_v". Fragmental level or conditional perplexity is not the only factor that affects or relates to ASR performance. For example, the landmark of a phoneme, which can sometimes be segmental, also contributes to improving the ASR accuracy of the phoneme.

### *B. Linking Perplexity, Discriminability, and Categorization*

If a model can distinguish two phonemes, the distance between model representations of the phonemes should be large. We attempt to quantify such model discriminability by the KL divergence (also called KL distance or relative entropy).

For example, when the phoneme /k/ is assigned with clusters of "1,2,3" and /g/ is assigned with "11,12,13," the model can distinguish two phonemes. The cluster distributions of /k/ and /g/ do not overlap, indicating a huge KL divergence between the cluster distributions. Accordingly, excellent model discriminability is quantified by the huge KL divergence.

We define the model discriminability of the phoneme pair of /k/ and /g/ as the KL divergence between the conditional distribution of the model clusters given /k/ and the conditional distribution of model clusters given /g/. We define the model discriminability of the distinctive feature of a voiced stop as the average KL divergence of the phoneme pairs of /b/-/d/, /b/-/g/, and /d/-/g/. Furthermore, we explored the relationship between model discriminability and infant discriminability (Fig. 9).

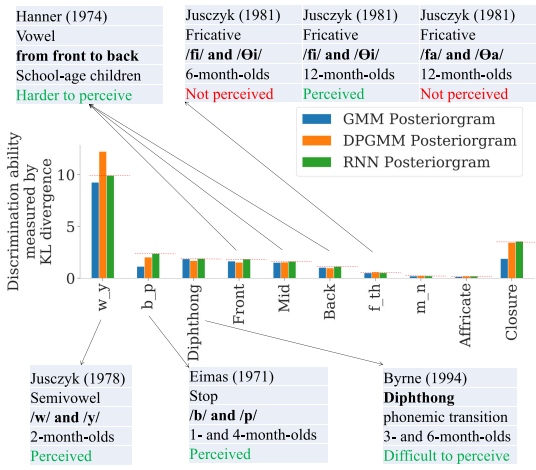



Fig. 9. Relation between the model discriminability and the infant discriminability of phonemes. We measured the model discriminability by the KL divergence (also called the KL distance or the relative entropy). The high discriminability corresponds to the large KL divergence. We define the KL divergence of a phoneme pair as the KL divergence between the conditional distribution of model clusters given the first phoneme and the conditional distribution of model clusters given the second phoneme. We define the KL divergence of a distinctive feature as the average KL divergence of all phoneme pairs that have the feature. We compute the conditional distributions by relative frequencies of the clusters.

We obtained the following observations from Fig. 9. The phonemes belonging to the semivowels, stops, and diphthongs that are distinguishable by infants have high KL divergence. The phonemes belonging to the fricatives and vowels that are not easily distinguishable by infants have low KL divergence. The discriminability of the DPGMM-RNN hybrid model is better than those of GMM and DPGMM.

Now we discuss the relationships among the perplexity, discriminability, and categorization (Fig. 10). Conditional perplexity describes the variability within a phoneme class. When a phoneme class corresponds to less fragmental cluster representations and fewer types of cluster segments, the conditional perplexity of the cluster distribution is small. KL divergence describes the variability between the phoneme classes. When two phonemes have little overlap in cluster distributions, the KL divergence between the cluster distributions is high. The model representation can achieve the phoneme categorization through low within-phoneme variability and high between-phoneme variability with low conditional perplexity and high KL divergence of the cluster distributions of phonemes.

In other words, a model categorizes phonemes well when the entropy (logarithm of perplexity) of the cluster distribution of a phoneme is low and the relative entropy (KL divergence) between cluster distributions of different phonemes is large.

## VI. Conclusion

We 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.

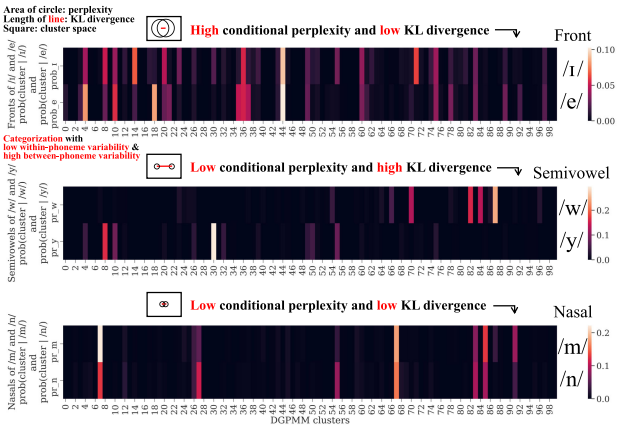



Fig. 10. Examples of cluster distributions of phoneme pairs that show the relations among conditional perplexity, KL divergence, and phoneme categorization. Conditional perplexity measures the uncertainty of the conditional distribution of the cluster given the phoneme. Low perplexity indicates that the distribution is concentrated on a few clusters. High perplexity indicates that the distribution is uniform over various clusters. The KL divergence measures the similarity between the conditional distributions. Distributions with small overlap have high KL divergence. The model categorizes the phonemes when the model cluster representations have low conditional perplexity and high KL divergence, where the representations show low variability within a phoneme and high variability across the phonemes. We compute the conditional distributions by the relative frequencies of the DPGMM clusters.

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.

## REFERENCES


[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE/ACM Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant Behav. Develop.*, vol. 7, no. 1, pp. 49–63, 1984.

[4] R. W. Semon, *The Mneme*. London, U.K.: George Allen and Unwin, 1921.

[5] B. Russell, *Analysis of Mind*. London, U.K.: George Allen and Unwin, 1921.

[6] S. A. Josselyn and S. Tonegawa, "Memory engrams: Recalling the past and imagining the future," *Science*, vol. 367, no. 6473, 2020.

[7] S. A. Josselyn, S. Köhler, and P. W. Frankland, "Heroes of the Engram," *J. Neurosci.*, vol. 37, no. 18, pp. 4647–4657, 2017.

[8] L. G. Reijmers, B. L. Perkins, N. Matsuo, and M. Mayford, "Localization of a stable neural correlate of associative memory," *Science*, vol. 317, no. 5842, pp. 1230–1233, 2007.

[9] X. Liu *et al.*, "Optogenetic stimulation of a hippocampal engram activates fear memory recall," *Nature*, vol. 484, no. 7394, pp. 381–385, 2012.

[10] J.-H. Han *et al.*, "Selective erasure of a fear memory," *Science*, vol. 323, no. 5920, pp. 1492–1496, 2009.

[11] S. Ramirez *et al.*, "Creating a false memory in the hippocampus," *Science*, vol. 341, no. 6144, pp. 387–391, 2013.

[12] G. Vetere *et al.*, "Memory formation in the absence of experience," *Nature Neurosci.*, vol. 22, no. 6, pp. 933–940, 2019.

[13] S. J. Martin, P. D. Grimwood, and R. G. Morris, "Synaptic plasticity and memory: An evaluation of the hypothesis," *Annu. Rev. Neurosci.*, vol. 23, no. 1, pp. 649–711, 2000.

[14] J. A. Kauer and R. C. Malenka, "Synaptic plasticity and addiction," *Nat. Rev. Neurosci.*, vol. 8, no. 11, pp. 844–858, 2007.

[15] W. G. Penfield, "Ferrier lecture - Some observations on the cerebral cortex of man," *Roy. Soc. London. Ser. B- Biol. Sci.*, vol. 134, no. 876, pp. 329–347, 1947.

[16] J. Locke, *An Essay Concerning Human Understanding*. London, U.K.: Thomas Basset, 1690.

[17] D. Hume, *An Enquiry Concerning Human Understanding*. London, U.K.: Andrew Millar, 1748.

[18] A. G. Samuel, "Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration," *Cogn. Psychol.*, vol. 88, pp. 88–114, 2016.

[19] P. D. Eimas and J. D. Corbit, "Selective adaptation of linguistic feature detectors," *Cogn. Psychol.*, vol. 4, no. 1, pp. 99–109, 1973.

[20] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," *Cogn. Psychol.*, vol. 47, no. 2, pp. 204–238, 2003.

[21] W. Penfield and L. Roberts, *Speech and Brain Mechanisms*. Princeton, NJ, USA: Princeton Univ. Press, 1959.

[22] J. H. McDermott, Audition. Oxford, UK: Oxford Univ. Press, 2014.

[23] J. M. Nielsen, *Agnosia, Apraxia, Aphasia: Their Value in Cerebral Localization*. New York, NY, USA: Paul B. Hoeber, Inc., 1946.

[24] S. Freud, *Zur Auffassung Der Aphasien: Eine Kritische Studie*. Leipzig, Germany: Franz Deuticke, 1891.

[25] S. Curtin, D. Hufnagle, K. Mulak, and P. Escudero, *Speech Perception: Development*. The Netherlands: Elsevier, 2017, pp. 1–7.

[26] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2000, pp. 1635–1638.

[27] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTERSPEECH*, 2010, pp. 1045–1048.

[28] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, pp. 209–230, 1973.

[29] M. Heck, S. Sakti, and S. Nakamura, "Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to Zerospeech 2017," in *Proc. Autom. Speech Recognit. Understanding*, 2017, pp. 740–746.

[30] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *Proc. INTERSPEECH*, 2015, pp. 3189–3193.

[31] S. Feng, T. Lee, and Z. Peng, "Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling," in *INTERSPEECH*, 2019, pp. 1093–1097.

[32] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proc. INTERSPEECH*, 2013, pp. 1781–1785.

[33] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge," in *Proc. INTERSPEECH*, 2015, pp. 3199–3203.

[34] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proc. INTERSPEECH*, 2015, pp. 3179–3183.

[35] A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "VQVAE unsupervised unit discovery and multi-scale Code2Spec inverter for zerospeech challenge," in *INTERSPEECH*, 2019, pp. 1118–1122.

[36] C. Manenti, T. Pellegrini, and J. Pinquier, "Unsupervised speech unit discovery using k-means and neural networks," in *Proc. Int. Conf. Stat. Lang. Speech Process.*, 2017, pp. 169–180.

[37] L. Ondel, L. Burget, and J. Černocký, "Variational inference for acoustic unit discovery," *Procedia Comput. Sci.*, vol. 81, pp. 80–86, 2016.

[38] J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj, "Hidden Markov model variational Autoencoder for acoustic unit discovery," in *Proc. INTERSPEECH*, 2017, pp. 488–492.

[39] C.-Y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. Assoc. Comput. Linguistics*, 2012, pp. 40–49.

[40] B. Wu, S. Sakti, and S. Nakamura, "Incorporating discriminative DPGMM posteriorgrams for low-resource ASR," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 201–208.

[41] S. Goldwater, M. Johnson, and T. L. Griffiths, "Interpolating between types and tokens by estimating power-law generators," in *Proc. Neural Inf. Process. Syst.*, 2006, pp. 459–466.

[42] N. Feldman, T. Griffiths, and J. Morgan, "Learning phonetic categories by learning a lexicon," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 31, 2009, pp. 2208–2213.

[43] N. H. Feldman, E. B. Myers, K. S. White, T. L. Griffiths, and J. L. Morgan, "Word-level information influences phonetic learning in adults and infants," *Cognition*, vol. 127, no. 3, pp. 427–438, 2013.

[44] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, no. 3, pp. B101–B111, 2002.

[45] B. De Boer and P. K. Kuhl, "Investigating the role of infant-directed speech with a computer model," *Acoust. Res. Lett. Online*, vol. 4, no. 4, pp. 129–134, 2003.

[46] B. McMurray, R. N. Aslin, and J. C. Toscano, "Statistical learning of phonetic categories: Insights from a computational approach," *Devlop. Sci.*, vol. 12, no. 3, pp. 369–378, 2009.

[47] D. Görür and C. E. Rasmussen, "Dirichlet process Gaussian mixture models: Choice of the base distribution," *J. Comput. Sci. Technol.*, vol. 25, no. 4, pp. 653–664, 2010.

[48] Y. W. Teh, "Dirichlet process," *Encyclopedia Mach. Learn.*, vol. 1063, pp. 280–287, 2010.

[49] B. Wu, S. Sakti, J. Zhang, and S. Nakamura, "Tackling perception bias in unsupervised phoneme discovery using DPGMM-RNN hybrid model and functional load," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 348–362, 2021, doi: 10.1109/TASLP.2020.3042016.

[50] B. Wu, S. Sakti, J. Zhang, and S. Nakamura, "Optimizing DPGMM clustering in zero-resource setting based on functional load," in *SLTU*, vol. 1, 2018, pp. 1–5.

[51] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1999.

[52] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 410–420.

[53] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI, Tech. Rep. 93, 1993.

[54] D. B. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *ICSLP*, 1992, pp. 899–902.

[55] P. Godard *et al.*, "A very low resource language speech corpus for computational language documentation experiments," *Comput. Res. Repository*, vol. abs/1710.03501, 2017. [Online]. Available: https://github.com/besacier/mboshi-french-parallel-corpus

[56] A. Bills *et al.*, "IARPA Babel Javanese Language Pack IARPA-babel402b-v1.0b LDC2020S07," *Linguistic Data Consortium*, 2020.

[57] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. Autom. Speech Recognit. Understanding*, 2011, pp. 1–4.

[58] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[59] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," Univ. British Columbia, Tech. Rep., vol. 1, 2007.

[60] R. P. Beapami, R. Chatfield, G. Kouarata, and A. Waldschmidt, *Dictionnaire Mbochi-Français*. Point Noire, Congo: SIL-Congo, 2000.

[61] L. Bouquiaux and J. M. Thomas, *Enquête Et Description Des Langues à Tradition Orale*, vol. 1. Leuven, Belgium: Peeters Publishers, 1976.

[62] J. Cooper-Leavitt, L. Lamel, A. Rialland, M. Adda-Decker, and G. Adda, "Corpus base linguistic exploration via forced alignments with a light-weight ASR tool," in *Proc. Lang. Technol. Conf., Hum. Lang. Technol. Challenge Comput. Sci. Linguistics*, 2017.

[63] M. Heck, S. Sakti, and S. Nakamura, "Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero-resource scenario," *Procedia Comput. Sci.*, vol. 81, pp. 73–79, 2016.

[64] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4960–4964.

[65] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP*, 2015, pp. 1412–1421.

[66] B. Van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the Zerospeech 2020 challenge," in *INTERSPEECH*, 2020, pp. 4836–4840.

[67] K. N. Stevens, *Acoustic Phonetics*, vol. 30. Cambridge, MA, USA: MIT Press, 2000.

[68] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 4835–4839.

[69] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 976–989, 2020, doi: 10.1109/TASLP.2020.2977776.

[70] M. Versteegh *et al.*, "The Zero Resource Speech Challenge 2015," in *Proc. INTERSPEECH*, 2015, pp. 3169–3173.

[71] E. Dunbar *et al.*, "The Zero Resource Speech Challenge 2017," in *Proc. Autom. Speech Recognit. Understanding*, 2017, pp. 323–330.

[72] E. Dunbar *et al.*, "The zero resource speech challenge 2019: TTS without T," in *INTERSPEECH*, 2019, pp. 1088–1092.

[73] R. E. Eilers, W. R. Wilson, and J. M. Moore, "Developmental changes in speech discrimination in infants," *J. Speech Hear. Res.*, vol. 20, no. 4, pp. 766–780, 1977.

[74] P. D. Eimas, E. R. Siqueland, P. Jusczyk, and J. Vigorito, "Speech perception in infants," *Science*, vol. 171, no. 3968, pp. 303–306, 1971.

[75] J. Bertoncini, R. Bijeljac-Babic, S. E. Blumstein, and J. Mehler, "Discrimination in neonates of very short CVs," *J. Acoust. Soc. Amer.*, vol. 82, no. 1, pp. 31–37, 1987.

[76] S. E. Trehub, "Infants' sensitivity to vowel and tonal contrasts," *Devlop. Psychol.*, vol. 9, no. 1, pp. 91–96, 1973.

[77] P. J. Swoboda, P. A. Morse, and L. A. Leavitt, "Continuous vowel discrimination in normal and at risk infants," *Child Develop.*, vol. 47, pp. 459–465, 1976.

[78] P. W. Jusczyk, *Infant Speech Perception: A Critical Appraisal*, J. L. Miller and P. D. Eimas, Eds. New York, NY, USA: Psychol. Press, 1982, pp. 113–164.

[79] R. E. Eilers and D. K. Oller, "The role of speech discrimination in developmental sound substitutions," *J. Child Lang.*, vol. 3, no. 3, pp. 319–329, 1976.

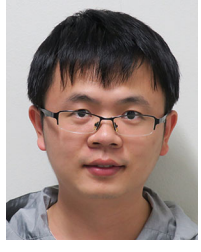
[80] M. S. Abbs and F. D. Minifie, "Effect of acoustic cues in fricatives on perceptual confusions in preschool children," *J. Acoust. Soc. Amer.*, vol. 46, no. 6B, pp. 1535–1542, 1969.

[81] M. A. Hanner, "Auditory discrimination and phonetic contexts in school age children," Master's thesis, Dept. Speech Pathology, Eastern Illinois Univ., 1974. [Online]. Available: https://thekeep.eiu.edu/theses/3625/

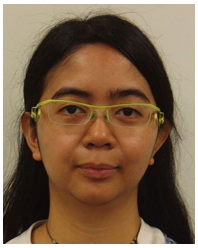
[82] P. W. Jusczyk, H. Copan, and E. Thompson, "Perception by 2-month-old infants of glide contrasts in multisyllabic utterances," *Percep. Psychophys.*, vol. 24, no. 6, pp. 515–520, 1978.

[83] J. M. Byrne, C. L. Miller, and B. Hondas, "Psychophysiologic and behavioral responsivity to temporal parameters of acoustic stimuli," *Infant Behav. Develop.*, vol. 17, no. 3, pp. 245–254, 1994.

[84] J. L. Miller and P. D. Eimas, "Studies on the perception of place and manner of articulation: A comparison of the labial-alveolar and nasal-stop distinctions," *J. Acoust. Soc. Amer.*, vol. 61, no. 3, pp. 835–845, 1977.

**Bin Wu** received the B.E. degree in computer science from the Jiangsu University of Science and Technology, Zhenjiang, China and the M.E. degree in computer applied technology from Beijing Language and Culture University, Beijing, China. He is currently working toward the Ph.D. degree with the Nara Institute of Science and Technology, Ikoma, Japan. His research interests include computational modeling of phoneme discovery and speech recognition.



**Sakriani Sakti** (Member, IEEE) received the B.E. degree in informatics (*cum laude*) from the Bandung Institute of Technology, Bandung, Indonesia, in 1999, the M.Sc. and Ph.D. degrees from the University of Ulm, Germany in 2008 and 2002, respectively. During her thesis work, she was with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003 and 2009, she was a Researcher with ATR SLC Labs, Japan, and during 2006–2011, she was an Expert researcher with NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study from 2005 to 2008, with Dialog Systems Group, University of Ulm. She was actively involved in collaboration activities such as Asian Pacific Telecommunity Project from 2003 to 2007, A-STAR, and U-STAR from 2006 to 2011). From 2009 to 2011, she was a Visiting Professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011 to 2017, she was an Assistant Professor with Augmented Human Communication Laboratory, NAIST, Japan. She was a Visiting Scientific Researcher of INRIA Paris-Rocquencourt, France, from 2015 to 2016, under JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation. From 2018 to 2021, she was a Research Associate Professor with NAIST and a Research Scientist with RIKEN, Center for Advanced Intelligent Project AIP, Japan. She is currently an Associate Professor with JAIST, an Adjunct Associate Professor with NAIST, a Visiting Research Scientist with RIKEN AIP, and an Adjunct Professor with the University of Indonesia. Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition and synthesis, spoken language translation, affective dialog system, and cognitive-communication. In 2000, she was the recipient of the DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany. She is a Member of JNS, SFN, ASJ, ISCA, IEICE, and IEEE. She is also a Committee Member of IEEE SLTC from 2021 to 2023, and an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2020 to 2023. Furthermore, she is the Chair of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a Board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU).

**Jinsong Zhang** (Member, IEEE) received the B.E. degree in electronic engineering from the Hefei University of Technology, Hefei, China, in 1989, the M.E. degree in electronic circuit, signal and system from the University of Science and Technology of China (USTC), Hefei, China, in 1992, and the Ph.D. degree in information and communication engineering from the University of Tokyo, Tokyo, Japan in 2000. From 1992 to 1996 he was a Teaching Assistant and a Lecturer with the Department of Electronic Engineering, USTC. From 2000 to 2007, he was an invited and Senior Researcher with the ATR Spoken Language Translation Research Laboratories. He is currently a Professor with the School of Computer Sciences, Beijing Language and Culture University, Beijing, China. His research interests include speech recognition, prosody information processing, 2nd language acquisition, computer assisted pronunciation training.

**Satoshi Nakamura** (Fellow, IEEE) received the B.S. degree from the Kyoto Institute of Technology, Kyoto, Japan, in 1981, and the Ph.D. degree from Kyoto University, Kyoto, Japan, in 1992. He is currently a Professor of the Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Japan, and the Honorary Prof. of Karlsruhe Institute of Technology, Karlsruhe, Germany. He was an Associate Professor of the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, from 1994 to 2000 and the Director of the ATR Spoken Language Communication Research Laboratories from 2000 to 2008, and the Vice President of ATR from 2007 to 2008. He was the Director General of Keihanna Research Laboratories and the Executive Director of the Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan, from 2009 to 2010, and a Project Leader of the Tourism Information Analytics Team, Center for Advanced Intelligence Project AIP of RIKEN Institute from 2017 to 2021. He is currently the Director of the Augmented Human Communication Laboratory and a Full Professor of the Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include modeling and systems of speech-to-speech translation and speech recognition. He is one of the Leaders of speech-to-speech translation research and has been serving for various worldwide speech-to-speech translation research projects, including C-STAR, IWSLT, and A-STAR. He was the recipient of the LREC Antonio Zampolli Award in 2012, Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics, Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He was an Elected Board Member of the International Speech Communication Association, ISCA in 2011–2018, an *IEEE Signal Processing Magazine* Editorial Board Member in 2012–2014, an IEEE SPS Speech and Language Technical Committee Member in 2013–2015. He is ISCA Fellow, IPSJ Fellow, and ATR Fellow.