Tackling Perception Bias in Unsupervised Phoneme Discovery Using DPGMM-RNN Hybrid Model and Functional Load

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

***Abstract*—The human perception of phonemes is biased against speech sounds. The lack of correspondence between perceputal phonemes and acoustic signals forms a big challenge in designing unsupervised algorithms to distinguish phonemes from sound. We propose the DPGMM-RNN hybrid model that improves phoneme categorization by relieving the fragmentation problem. We also merge segments with low functional load, which is the work done by segment contrasts to differentiate between utterances, just like humans who convert unambiguous segments into phonemes as units for immediate perception. Our results show that the DPGMM-RNN hybrid model relieves the fragmentation problem and improves phoneme discriminability. The minimal functional load merge compresses a segment system, preserves information and keeps phoneme discriminability.**

***Index Terms*—Unsupervised phoneme discovery, perception of phonemes, DPGMM, RNN, functional load, zerospeech.**

I. Introduction

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

Unsupervised phoneme discovery [1], [2] or similar tasks [3]–[5] have been explored by different experiment settings with different measurements. Recently the Zero Resource Speech Challenge [6] was organized to compare the performance of these methods. Typical methods include neural network technology, such as representation learning by autoencoder [7]–[9] or discriminative training by ABnet [10], traditional clustering such as GMM [11] or k-means [11], [12], and nonparametric clustering such as the Dirichlet Process Gaussian Mixture Model (DPGMM) trained by Gibbs sampling [1], or variational inference [13], [14]. Among them, DPGMM, which is acoustic clustering, achieved the top performance at Zerospeech 2015 and 2017 [15], [16].

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

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

A. Bias of Phoneme Perception

Identifying phonemes from natural speech is challenging. Early studies on the high correlation between sound spectra and isolated phonemes provided encouragement that the problem could be solved. For example, we can identify vowels by formant values or stops by silent periods, which are verified by the speech synthesis practice [19], [20]. However, seeking phonemes from the spectrum in spontaneous speech flow is frustrating. While we are listening to some phonemes, features, or breaks at certain moments, we cannot find enough evidence about them from the spectrum [21]. The spectrum faithfully reacts to energy of different frequencies at a certain moment; it doesn't react to the sound history or subsequent sounds. However, our phoneme perception is biased. Instead of merely momentarily decoding the speech, our perception is influenced by the expectation of what will come next or our speaking and hearing experience. The lack of correspondence between speech perception and sound stream forms a central challenge in phoneme discovery from spontaneous speech [22].

The human perception of phonemes is biased against speech sounds. For example, when a virtually identical burst happens before /i/, /a/, or /u/, we tend to hear /pi/, /ka/, or /pu/ [21] because we hear them while referencing how we say them [22]. Since the

Manuscript received March 18, 2020; revised July 23, 2020 and September 25, 2020; accepted November 8, 2020. Date of publication December 2, 2020; date of current version December 18, 2020. *(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).

Sakriani Sakti and Satoshi Nakamura are 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@is.naist.jp; s-nakamura@is.naist.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.2020.3042016

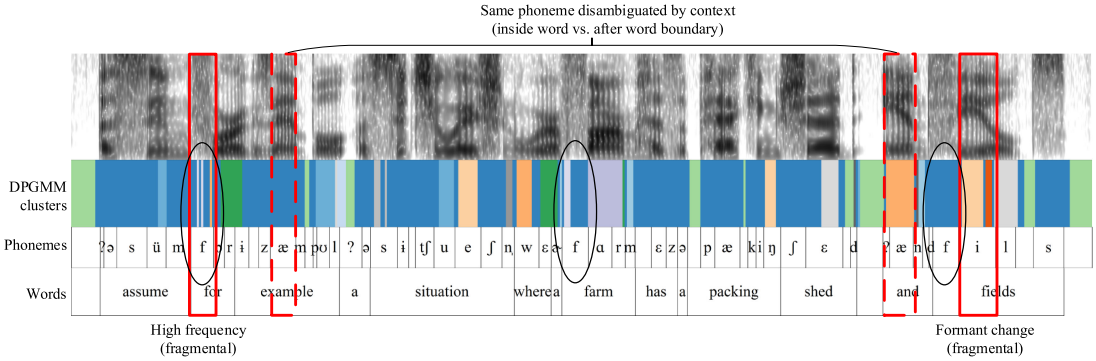


Fig. 1. Example [17] that shows problems of DPGMM clustering in unsupervised phoneme discovery from TIMIT corpus [18]. Top layer is spectrogram followed by DPGMM label, phoneme, and word layers. In second layer, each color denotes one specific type of DPGMM cluster. Red, solid-lined rectangles show that complex acoustics such as fricatives with high frequency and vowels with rapid formant change cause fragmental DPGMM clusters; small black circles show that we can improve categorization of identical phoneme if there are no fragments; red, dash-lined rectangles show two acoustically different segments disambiguated by surroundings, which should be same phoneme, are treated as different clusters.

lips are close together when we generate /i/ and /u/, this bilabial articulation may interpret the burst as /p/; when the tongue is relatively low and back when we generate /a/, dorso-velar articulation may interpret the burst as /k/. Visual context can also bias our perception [23]. When hearing a recording /ba/ while watching a video of a face saying /ga/, the listeners report that they hear /da/. The compromise between visual and auditory information indicates that with accurate visual information, we can probably correct the phonemes.

Human perception has an "auto filling" ability for perceiving phonemes in sound streams [24]. Even when a phoneme (with its transition cues) is replaced by noise, people report they hear it and don't notice any noise or its location. Our lexicon knowledge influences our perception of phonemes. By adding an identical, intermediate sound between /d/ and /t/ in front of "ask" and "ash," Ganong found that people reported hearing "task" rather than "dask" and "dash" rather than "tash" [25]. Sometimes our perception relies less on lexical knowledge and more on the probability of sequences of phonemes (e.g., compensation for co-articulation varies with phonotactic probability [26]). A person's speaking and listening experiences, including the segment probabilities or sequences as well as how he says or hears these segments to achieve economical communication, also implicitly bias his perception.

Phoneme perception categorizes acoustic sounds [22], which shows another fundamental perceptual bias. If we create linearly changed acoustic stimuli between two phonemes, such as /t/ and /d/, our perception nonlinearly jumps from one category to another because we cannot identify different acoustic realizations inside one phoneme category.

The above studies show that our phoneme perception is biased. Perception bias becomes a big problem in unsupervised phoneme discovery (Zero Resource Speech Challenge [27], as we introduced at the beginning of our paper), which asks machines to learn phonemes from acoustic speech in an unsupervised way [1], [27]. A machine learning algorithm discovers *objective* acoustic segments from speech, while humans annotate *subjective* perceptual phonemes as underground truth with perception bias. For example, in Fig. 1, the clustering algorithm treats the same phoneme /æ/ in 'example' and 'and' as different acoustic segments because their acoustical spectra are quite different; it treats the same phoneme /f/ in 'for,' 'farm,' and 'fields' as different acoustic segments by faithfully recording the acoustic fragmental realizations inside the phoneme category.

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

B. DPGMM-RNN Model and Phoneme Categorization

Machines can directly get discrete segments by applying such clustering algorithms as K-means [11], [12], GMM [11], or DPGMM clustering [1], [13], [14] from the acoustic features. The DPGMM algorithm [28] retained the state-of-the-art approach in the Zerospeech 2015 and 2017 [15], [16].

However, framewisely clustering acoustic features to get segments suffers from the intra-phoneme fragmentation problem (Fig. 1). First, these traditional clustering algorithms cannot fully capture the temporal information of speech features. As long as the spatial distribution of these acoustic features in high-dimensional space does not change, such clustering algorithms as K-means or GMM always get similar results because they ignore the time order of these features. The DPGMM algorithm introduces the Dirichlet Process (DP) to help dynamically create new clusters at every moment based on the frequency of the clusters of all the previous frames without considering their order [29]. Theoretically, DP is infinitely exchangeable; joint distribution doesn't depend on the order of data if they are infinite [30]. We believe DPGMM involves weak temporal contextual modeling for finite sequential data clustering. Second, in actual unsupervised phoneme discovery practices, after carefully tuning the parameters (e.g., DPGMM's concentration parameter, which is closely related to the number of clusters), such optimal performances (in discriminating phonemes in different languages) always create more clusters than the number of phonemes in normal human languages [16], [17]. Third, the

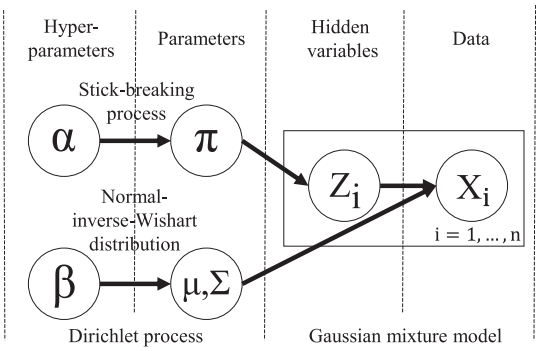


Fig. 2. Graphical model of Dirichlet Process Gaussian Mixture Model (DPGMM). We generate parameters of weights ($\pi = \pi_1, \ldots, \pi_k, \ldots$), means, and variances ($(\mu, \Sigma) = (\mu_1, \Sigma_1), \ldots, (\mu_k, \Sigma_k), \ldots$) for Gaussians from stick-breaking process (with concentration parameter α) and normal-inverse-Wishart distribution (with parameter $\beta = (\mu_0, \lambda, \Sigma_0, \nu)$) respectively. We then generate each frame of speech feature X_i of data $X = X_1, \ldots, X_n$ by first sampling one Gaussian with mean μ_k and variance Σ_k indicated by hidden variable $Z_i = k$ according to weights, and sampling X_i from that Gaussian cluster. The box, with (Z_i, X_i) inside, is a simplified notation of all n data points (features) with their indicator hidden variables $((Z_1, X_1), \ldots, (Z_i, X_i), \ldots, (Z_n, X_n))$.

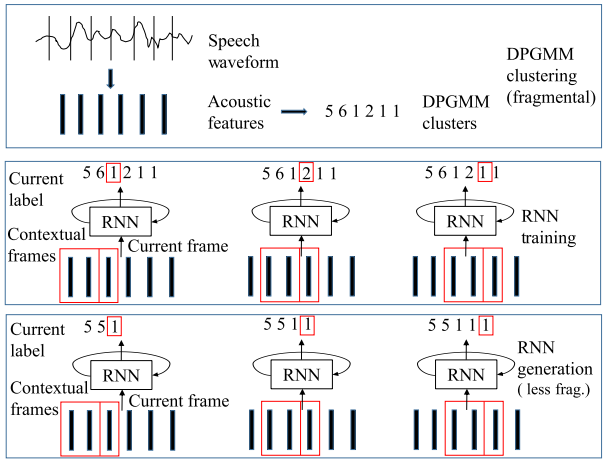


Fig. 3. Three steps to construct DPGMM-RNN hybrid model for unsupervised phoneme discovery: using RNN to relieve segmentation problem of DPGMM clusters.

DPGMM algorithm creates small clusters inside one phoneme—the fragmentation problem—with such a complex acoustic structure as a fricative with high-frequency components or a vowel with a sharp format change [17]. The DPGMM algorithm tries to get higher resolution by struggling to discriminate among acoustically complex phonemes, which also tends to increase the number of clusters overall.

Human perception of speech is categorical [31]. We don't hear intra-phoneme fragmental details when discriminating complex phonemes [24], [31]. For general sounds, for example, people can discriminate about 2000 different pitches, but they can only identify about seven absolute ones [22]. For speech sounds, however, similar discriminability and identifiability of phonemes make people fail to discriminate the acoustic variations inside each phoneme category [31]. If we believe one phoneme type is a set of segments, then our biased perception cannot distinguish within the set, including the unstable fragmental acoustic realizations, created from the clustering algorithm, of these segments.

In this paper, we propose the DPGMM-RNN hybrid model, which uses RNN to relieve the DPGMM intra-phoneme fragmentation problem. Assume that human perception is hierarchical at low-level perception. At the first run, very low-level, bottom-up unbiased clustering can get fragmental details with sufficient discriminability of the segments from the raw stimuli of speech by air vibration. At the second run, the ear uses a primitive top-down acoustical contextual refinement and pays little attention to the variations inside one phoneme. Such perceptual refinement can be achieved by RNN mapping (Fig. 3). We train RNN intensively by remembering the phoneme (DPGMM clusters) at every moment of speech by listening to a chunk of sound that includes that moment. Listening by chunks helps integrate long acoustical contexts as a whole instead of concentrating on random short-time fragmental changes over a few frames. After RNN remembers different chunk realizations for each phoneme, it has the ability to identify the sounds. We show that RNN refinement relieves the fragmentation problem inside phoneme categories.

Moreover, facing the weak contextual modeling of DPGMM, whose joint likelihood does not depend on the order of the observed data when they are infinite [30] and mainly captures acoustic information at the frame level, RNN refinement with chunks of successive frames instead of single frames of speech can rediscover such temporal structures as formant transitions, which cross several frames and are important acoustic cues to perceptually discriminate phonemes in spontaneous speech [22].

In human perception, top-down contextual constraints, not only acoustical or phonemic ones but also linguistic and lexical ones help correct or remove the segments that are wrongly discovered to make them closer to phoneme units [25]. In previous research [32]–[34] on infant phoneme acquisition, the words and phonemes are assumed to be acquired at the same period and jointly optimize each other, supported by phoneme-lexicon joint discovery by adaptor grammar framework [35] and hierarchical nonparametric Bayesian model [36]. Language model of discovered segments was trained to optimize phoneme discovery [37]. In our proposed DPGMM-RNN hybrid model, RNN remembers the statistical structure, reflecting on such contextual constraints at the phonetic and lexicon levels, of audio segments from DPGMM clusters. For example, if the DPGMM clustering algorithm confuses 'bite' with 'kite' inside an utterance that contains the sound 'dog,' RNN can correct such mistakes because sounds 'dog' and 'bite' are semantically correlated. RNN remembers their co-occurrence.

C. Functional Load and Economical Principle

Phonemes are not the only discrete perceptual categories decoded from acoustic implementations. They also compose the hierarchical linguistic structure that represents knowledge for communication. From the view of economical communication, we want to maintain the information transmission rate and simultaneously reduce the structure of the phoneme sequence and the size of the phoneme system.

People communicate economically by perceptually identifying only a few types of sound patterns (phonemes) and convey complex ideas with a few tokens of linguistic or perceptual units during a single utterance (low bit rate). An economical phoneme system is required by the physiological constraints of the human auditory system. First, we can only use a few segments to quickly convey information. For example, if you listen to a YouTube video of recorded speech and increase its speed fourfold, hearing all of the linguistic units or completely comprehending the speech is almost impossible. More than 30 phonemes per second exceeds the normal temporal resolving power of the human ear [38]. Second, the normal human ear has limited capability to identify sounds, usually considerably fewer than 40 sounds (for segments under 50 milliseconds in casual conversation) when people communicate rapidly in daily life [39]. When designing algorithms for phoneme discovery, just considering the acoustics may be inadequate, and such knowledge as global frequency or information load over the segments becomes important for optimizing the high information rate constrained by the human ear's listening ability.

Unsupervised phoneme discovery needs economical representation from continuous speech while correctly discriminating among phonemes [27] (with low bit rate and low discriminative error). Removing such non-linguistic information as speakers and stress increases economical representation. Mean and Variance Normalization (MVN) and Vocal Tract Length Normalization (VTLN) can normalize speaker information [16], and speaker-adversarial learning [40] creates speaker independent features [41]. Maximum Likelihood Linear Transform (MLLT) and Speaker Adaptive Training (SAT) deal with the speaker variability of acoustic features [42]. Economical representation comes from compression. PCA and LDA compress the acoustic features [43], and auto-encoder reduces their dimensions [8]. However, low-dimensional continuous features are never as efficient as discrete features or discrete segments. The Vector Quantised-Variational AutoEncoder (VQ-VAE) can quantize speech acoustic features [11].

In this paper, we design a phoneme system that is economical over the language's structure because a phoneme system's function not only acoustically discriminates the phonemes themselves but more importantly distinguishes the language utterances [21]. For example, we get such a discrete segment sequence as 'A B C 12 13 14 D 12 13 14 E 12' from speech. The simplest local structure is center segments with first order surroundings: <s >-A-B A-B-C B-C-12 ... 12-13-14 ... 12-13-14 ... 14-E-12 E-12-< s>. For this toy language, segment 'B' is always surrounded as 'A-B-C'; segment '13' is always surrounded as '12-13-14'. We describe that segments 'B' and '13' have a complementary distribution [21]. If we design a phoneme system for this toy language, then 'B' and '13' should be the same phoneme, because they are disambiguated by ('A', 'C') and ('12', '14').

We propose to merge the segments with low functional load [17], [44] , which is defined as the segment contrasts' work to differentiate utterances [21], [45]–[47]. Segments with low functional load can usually be disambiguated by their surroundings and convey little information [45].

Merging segments with low functional load follows the general economical principle of speech communication [48], which argues that as long as we can convey the essential information, we don't focus on clearly hearing each sound or its acoustic details.

We propose the DPGMM-RNN hybrid model for decoding segments from speech signals. The DPGMM algorithm finds fragmental segments, while RNN fixes the fragmentation problem. We also merge unambiguous segments with small functional load to get a more efficient segment system.

From the view of using machine learning to track the bias of human perception, our DPGMM-RNN hybrid model achieves better phoneme categorization by solving the fragmentation problem. Merging segments with low functional load just like done by humans turns unambiguous segments into phonemes and qualifies them as units for immediate perception [22].

In our evaluation of our proposals, we use conditional entropy, which is the average number of clusters per phoneme for measuring the fragmental level of the discovered segments. We also use the ABX discriminability score [49], which is the cluster represenation's ability in discriminating among phoneme.

II. Method

A. DPGMM-RNN Hybrid Model: Relieve Fragmentation Problem

The DPGMM-RNN hybrid model combines DPGMM clustering and RNN reconstruction. Let us first briefly introduce the former.

1) DPGMM Clustering Algorithm: We can treat DPGMM as an infinite GMM with density function $p(x_i) = \sum_{k=1}^{\infty} \pi_k p(x_i|\mu_k, \Sigma_k)$ (alternatively, $p(x_i) = \sum_{k=1}^{\infty} p(Z_i = k)p(x_i|Z_i = k)$).

This generative model (Fig. 2) is defined by the following procedures. It samples mixture weights $\{\pi_k\}_{k=1}^{\infty}$ from the stick-breaking process [50] (with concentration parameter α) and the means and variances $\{\mu_k, \Sigma_k\}_{k=1}^{\infty}$ from the normal-inverse-Wishart (NIW) distribution (with a belief of mean μ_0, the belief of variance Σ_0, the belief-strength of mean λ, and the belief-strength of variance ν). It also samples Gaussian cluster indicator hidden variable Z_i by mixture weights and each data point X_i by the Gaussian cluster indicated by Z_i. We summarize this sampling procedure by describing the dependency relation of the random variables of the joint distribution of model $\mathrm{DPGMM}(\alpha, \mathrm{NIW}(\mu_0, \lambda, \Sigma_0, \nu))$ in Fig. 2.

Given the model definition and data $\{x_i\}_{i=1}^{n}$, we can infer from the Gibbs sampling to get posterior $p(z_i|x_i)$ and the cluster of any data point x_i by $k^* = \mathrm{argmax}_k\, p(z_i = k|x_i)$.

2) DPGMM-RNN Hybrid Model: We generally construct the DPGMM-RNN hybrid model using RNN to refine the DPGMM clusters. We apply the hybrid model to the unsupervised phoneme discovery, which uses RNN to relieve the fragmentation problem of the DPGMM clusters in three steps (Fig. 3):

- **DPGMM clustering**: after extracting the features from the raw audio, we apply the DPGMM clustering algorithm to get a cluster label for each feature frame. Many DPGMM

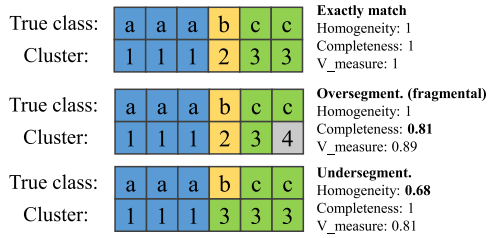


Fig. 4. Relationship between evaluation metrics (homogeneity, completeness, and v-measure) and mismatching problems (fragmentation, oversegmentation, and undersegmentation). All metrics range between 0 to 1; higher value means better matching.

segments (successive frames with identical cluster labels) are fragmental, which are much shorter (one or few frames) than phonemes in human language (Fig. 1). We use RNN to relieve this fragmentation problem.

- **RNN training**: we train the RNN model by mapping from a feature segment to the DPGMM cluster label of the last frame of that segment. RNN uses a shared memory to remember the tendency that momentarily produced the DPGMM cluster label from the nearest acoustic 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 reconstruction of cluster labels helps relieve the fragmentation problem (Fig. 5). For example, DPGMM fragmental structure "a a b a a" in several successive frames is usually revised by RNN to "a a a a a."

B. Minimal Functional Load Merge: Compress Segment System

Assume that the speech segments of a language are a discrete label sequence, and our goal is to compress the language's label set.

We merge the labels that can be disambiguated by their surroundings by brute-forcely recording the surroundings for each label. Another equivalent way [45] is to merge the labels that bear the minimal functional load [44], [51].

1) Theory of Functional Load: The functional load of a set of labels is computed by the loss of the language's entropy after merging these labels [51]:

$$FL(\alpha) = \frac{H(L) - H(L_\alpha)}{H(L)}, \tag{1}$$

where α is the set of labels, $H(L)$ is the entropy of language L, and $H(L_\alpha)$ is the language's entropy after α is merged. The entropy of the language is computed by the relative frequency of the strings of labels:

$$H(L) = -\frac{1}{K}\sum_{i=1}^{n} p(s_i)\log p(s_i), \tag{2}$$

where s_i is a string of labels and K is the length of the strings (See Appendix for an example of the computation of functional load).

Algorithm 1: Compress With Minimal Functional Load.

while the label set exceeds a threshold **do**

1) **Functional load calculation:** compute the functional load of each pair of labels for the language.
2) **Merge decision:** merge the label pair with the minimum functional load.

$$(x^*, y^*) = \arg\min_{(x,y)} FL(x, y) \tag{3}$$

3) **Update:** renew the label sequence (the language) by merging optimal label pair (x^*, y^*) by treating x^* and y^* as indential label.

end while

2) Compression with Minimal Functional Load: We find a set of labels to merge by greedily merging pairs with minimal functional load [44] (Algorithm 1).

III. Evaluation Metric

We evaluated our generated segments with conditional entropy-based measurements (conditional perplexity, homogeneity, completeness, and v-measure) and psychology-based measurements (ABX discriminability score and ABX error rate).

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

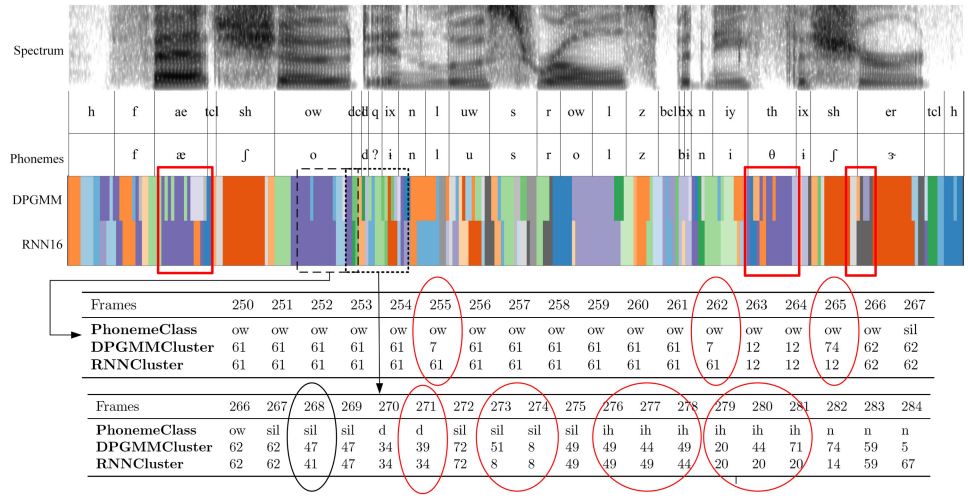
$$ppl(C|T) = 2^{H(C|T)}, \tag{4}$$

$$\begin{aligned} H(C|T) &= \sum_t p(t)H(C|T=t) \\ &= -\sum_t p(t)\sum_c p(c|t)\cdot logp(c|t) \\ &= -\sum_t \frac{n_t}{n}\sum_c \frac{n_{ct}}{n_t}\cdot log\frac{n_{ct}}{n_t}, \end{aligned} \tag{5}$$

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

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

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



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

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

Fig. 5. Utterance 'Fat showed in loose rolls beneath the shirt' with id FADG0_SI1909 from TIMIT test set to show neural network helps solve fragmentation problems. Top layer is spectrogram followed by phoneme layer, DPGMM, and RNN layers. RNN16 is short for DPGMM-RNN hybrid model with 16 contextual frames. Red rectangles show neural network reduces fragmental problem; table shows more details of agreement of classes and clusters in two segments (red circles for better cases in relieving fragmentation problem, black ones for worse cases).

need the amount of phoneme classes per cluster as an additional measurement.

Completeness, **homogeneity**, and **v-measure** (Fig. 4) are measurements [52] (similar to accuracy, recall, and F-score) that reflect the matching degree between generated clusters and underground phonemes using normalized conditional entropy. Completeness c, homogeneity h, and v-measure v (harmonic mean of h and c) are defined as follows:

$$c = 1 - \frac{H(C|T)}{H(C)}, \tag{6}$$

$$h = 1 - \frac{H(T|C)}{H(T)}, \tag{7}$$

$$v = \frac{c \cdot h}{c + h}. \tag{8}$$

We compute the entropy and the conditional entropy by the relative frequency, similar to Eq. (5) of the framewise samples from generated cluster C and phoneme truth T. All three measurements were normalized between 0 to 1 (as $H(T) \geq H(T|C)$), and the higher value shows better matching between the generated clusters and the underground phonemes. High completeness shows that each phoneme type almost 'completely' (completeness) corresponds to a unique generated cluster type; high homogeneity shows that one cluster type should correspond to the 'same' (homogeneity) phoneme truth type.

As shown in Fig. 4, low completeness indicates that the cluster representation is fragmental or oversegmental with respect to the phoneme truth. Low homogeneity indicates undersegmental. Only high v-measure indicates that the representation is neither oversegmental (fragmental) nor undersegmental because the clusters agree with the phonemes.

In addition to the above conditional entropy-based measurements, which are based on the global frequency, we also evaluate our representation by the discriminability of the local phoneme segments using psychological measurements: ABX discriminability score (or its reverse: ABX error rate) [49].

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

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

In theory, we can also define the discriminability score between the triphone category pair $(c1, c2)$ [53] by taking samples A and X from $c1$ and sample B from $c2$ and define score s and its point estimator as follows:

$$\begin{aligned} s(c1, c2) &= p(d(A, X) < d(B, X) | A, X \in c1, B \in c2) \\ &\quad + \frac{1}{2} p(d(A, X) = d(B, X) | A, X \in c1, B \in c2) \\ &= \frac{1}{m(m-1)n} \sum_{a \in c1} \sum_{x \in (c1 - \{a\})} \sum_{b \in c2} \\ &\quad (\delta_{d(a,x) < d(b,x)} + \frac{1}{2} \delta_{d(a,x) = d(b,x)}), \end{aligned} \tag{9}$$

where δ_c is the indicator function (taking value 1 if condition c is true and 0 if c is false). Coefficient m is the number of triphones of the $c1$ category, and n is the number of triphones of the $c2$ category. Metric d is any distance of the cluster representation between triphone segments, which are extracted by the phoneme annotation. We computed three specific distances between the triphone segments, with possible different number of frames, by Dynamic Time Warping (DTW) based on a frame-to-frame cosine distance (cos), symmetric Kullback-Leibler divergence (kl), and edit distance (edit) [53].

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

$$d_{cos}(x, y) = \frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{D} x_i y_i}{\sqrt{\sum_{i=1}^{D} x_i^2}\sqrt{\sum_{i=1}^{D} y_i^2}}, \tag{10}$$

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

$$\begin{aligned} d_{kl}(x, y) &= \frac{1}{2}KL(x||y) + \frac{1}{2}KL(y||x) \\ &= \frac{1}{2}\sum_{i=1}^{D} x_i log\frac{x_i}{y_i} + \frac{1}{2}\sum_{i=1}^{D} y_i log\frac{y_i}{x_i}, \end{aligned} \tag{11}$$

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

$$\begin{aligned} &d_{edit}(x, y) = d_{D,D}(x, y) \\ &d_{i,j}(x, y) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0, \\ \min \begin{cases} d_{i-1,j}(x, y) + 1 \\ d_{i,j-1}(x, y) + 1 \\ d_{i-1,j-1}(x, y) + \delta_{x_i \neq y_j} \end{cases} & \text{otherwise,} \end{cases} \end{aligned} \tag{12}$$

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

The **ABX error rate** [53] is defined as one minus the average of the discriminability scores of all the category pairs with corresponding triplets A, B and X. If (A, B, X) comes from the same speaker, we call it the ABX error rate within speakers. If (A, X) and (B, X) come from different speakers, then we call it the ABX error rate across speakers.

IV. Dataset and Experiment Setup

A. Dataset

We analyzed the DPGMM-RNN hybrid model with the test set of the TIMIT corpus [18], which contains 0.81 hours read speech with 1344 English utterances.

We compared the DPGMM-RNN hybrid model with the methods that achieved the top results in Zerospeech 2019 [27] with English read speech: 5941 training utterances spoken by 100 speakers (about 15 hours and 40 minutes) and 455 test utterances spoken by 24 speakers (about 28 minutes).

We tested the performance of the DPGMM clusters after minimum functional load compression using the Xitsonga corpus of African read speech with about 2 hours and 29 minutes of segments provided by Zerospeech 2015 [6].

B. Experiment Setup

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

We obtained clusters with the DPGMM algorithm using the same parameter setting as previous works [16], [43] with a toolkit [28]. We set the concentration parameter to 1 and the mean and variance of the prior to the global mean and the global variance of the MFCC features with belief-strengths 1 and 42. We got cluster labels after 1500 sampling iterations.

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

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

We got the conditional entropy-based measurements (conditional perplexity, completeness, homogeneity, and v-measure) by python and Scikit-learn [54]. We computed the ABX discriminability scores and the ABX error rates with a toolkit provided by Zerospeech 2015 and compared the DPGMM methods with other methods proposed in Zerospeech 2019 with its official evaluation program.

V. Result

A. DPGMM-RNN Hybrid Model and Fragmentation Problem

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

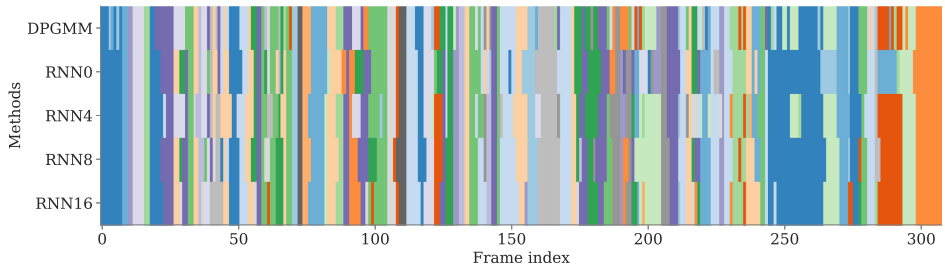


Fig. 6. Utterance example from TIMIT test set to show fragmental level of segments decreases by applying stronger contextual modeling of DPGMM-RNN hybrid model; RNNn denotes DPGMM-RNN hybrid model with n contextual frames as RNN input.

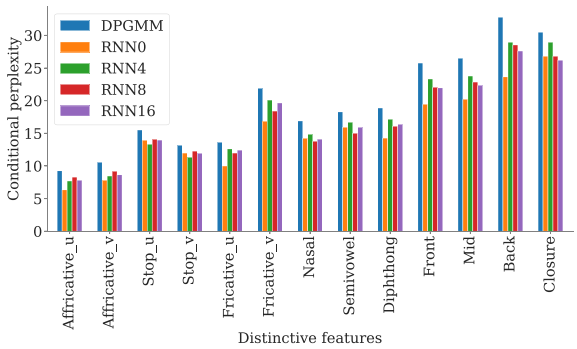
Fig. 5 shows that the DPGMM algorithm generates fragmental segments inside phonemes, and some fragments disappear after applying RNN reconstruction (as shown by red circles).

Fig. 6 shows that the DPGMM-RNN hybrid model decreases tiny fragments using RNN by accepting longer chunks, each of which is composed of the current frame and its past successive contextual frames, with stronger contextual modeling.

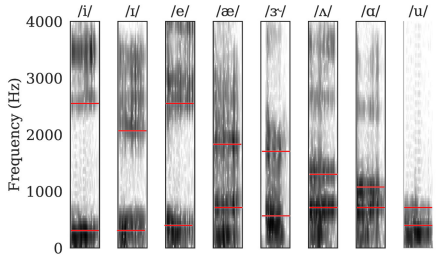
B. Distinctive Features and Fragmentation Problem

We explored why the fragmentation problem happens, how the sensitivity of the acoustics of the DPGMM clustering algorithm is associated with distinctive features, and how the proposed DPGMM-RNN hybrid model reacts to different distinctive features. After partitioning the set of phonemes into groups by distinctive features, we computed the conditional perplexity for each phoneme group to determine the average number of clusters per phoneme (the fragmental level) for each distinctive feature. Fig. 7(a) shows the following results.

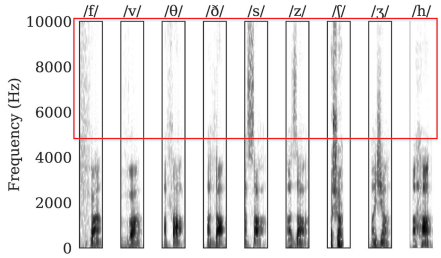
- The vowels are more fragmental than the consonants; voiced consonants are more fragmental than unvoiced ones.
- The vowels from front to back became more and more fragmental when the first and second formants became closer and harder to differentiate (Fig. 7(b)).
- The fricatives are the most fragmental consonants, which involve high-frequency components in the speech signals (Fig. 7(c)) and irregularity and rapid changes of acoustics.
- The DPGMM-RNN hybrid model (RNNn) can relieve the fragmentation problem of the DPGMM clusters (DPGMM) by decreasing the conditional perplexity for each distinctive feature.
- We computed the relative decrease ratio of the conditional perplexity between DPGMM and RNN16. Features that are more fragmental decreased more, except for affricatives and nasals.
- Even after applying RNN to relieve the fragmental problem, the conditional perplexity, which is the average number of clusters per phoneme, remained high for each feature, roughly between 5 to 20.



(a) Conditional perplexity of clusters given phoneme classes



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



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

Fig. 7. Upper subfigure (a): conditional perplexity to show fragmental level for each distinctive feature; RNNn denotes DPGMM-RNN hybrid model with n contextual frames. Middle subfigure (b): spectrogram of vowels from front to back; first and second formants are marked by red bars. Lower subfigure (c): Spectrogram of fricatives with high-frequency noisy components. We extended highest frequency from 4000 to 10000 Hz compared to subfigure (a) to see high-frequency components of fricatives (inside red rectangle).

C. Analysis of Cluster Agreement With Phoneme Class

1) Overall Performance: Fig. 8 shows that the DPGMM-RNN hybrid models (RNNLabel, RNNPost) outperformed the DPGMM algorithm (DPGMM) for homogeneity, completeness, and v_measure.

Since direct RNN learning from the discrete DPGMM label always gets better results than from the continuous DPGMM

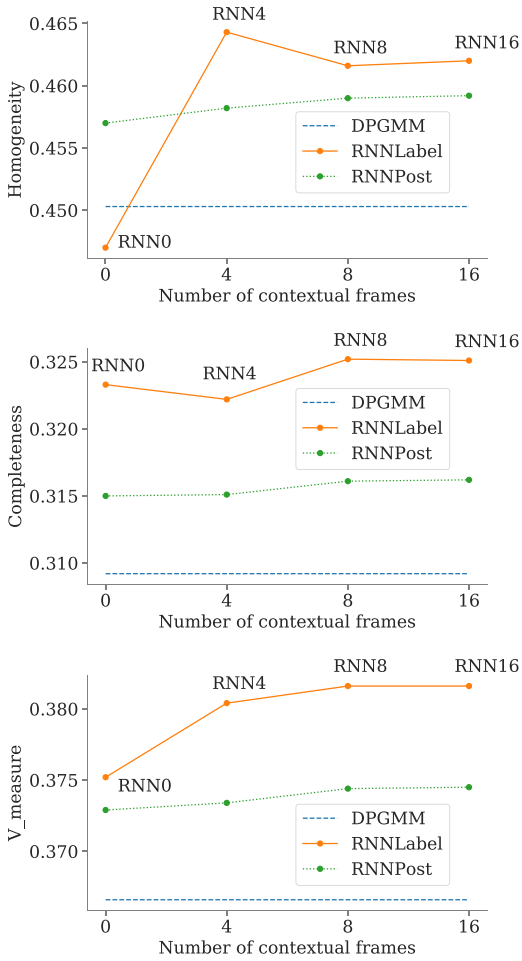


Fig. 8. Homogeneity, completeness, and v_measure scores of TIMIT test set to show matching degree between clusters and phonemes. Dashed line is DPGMM clustering scores, and solid and dotted lines are DPGMM-RNN hybrid model scores. RNNLabel learns from discrete DPGMM cluster label with cross entropy loss; RNNPost learns from the continuous DPGMM posteriorgram with MSE loss. RNNn denotes DPGMM-RNN hybrid model with n contextual frames.

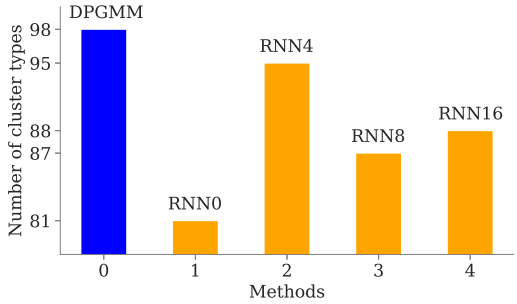


Fig. 9. Number of cluster types from DPGMM clustering (blue bar) and that from DPGMM-RNN hybrid models with 0, 4, 8, and 16 contextual frames (orange bars).

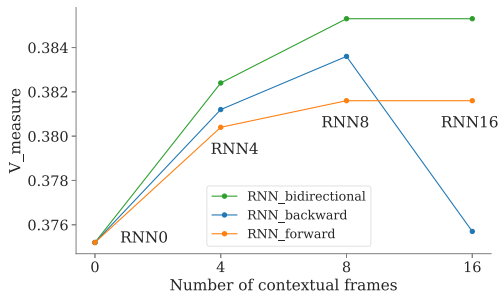


Fig. 10. v_measure of DPGMM-RNN hybrid model with different context models. For example, when using eight frames of acoustic features as RNN input context, RNN_forward takes a current frame along with eight past frames, RNN_bidirectional takes a current frame along with four past frames and four future frames, and RNN_backward takes a current frame and eight future frames.

posteriorgram, in later experiments, our hybrid model learned from the label (RNNn or RNNLabel) by default.

2) Contextual Modeling: As we increase the length of the context of the DPGMM-RNN hybrid model, the v_measure becomes larger (Fig. 8), showing better matching of the generated model clusters and the underground phoneme classes.

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

Fig. 10 shows that the hybrid model gains better v_measure by learning the RNN from both past and future acoustic features (RNN_bidirectional) compared to merely learning from the past (RNN_forward or RNNn) or the future (RNN_backward). Since the implementation of simpler models needs less effort and makes it easier for communities to reproduce our results, most DPGMM-RNN hybrid models of this paper used the simplest strategy: learning mapping from past acoustic features (RNNn).

3) Oversegmentation and Undersegmentation: RNN0 (the hybrid model without a contextual frame for RNN) has relatively low homogeneity and high completeness (Fig. 8), which suffers from the possible undersegmentation problem: the number of cluster types of RNN0 is lower than the others (Fig. 9). RNN4 (hybrid model with four contextual frames for RNN) has relatively high homogeneity and low completeness (Fig. 8), which suffers from the possible oversegmentation problem: the number of RNN4 cluster types is higher than the other hybrid models (Fig. 9). Both homogeneity and completeness increase from RNN8 to RNN16.

4) Representation Compression: Fig. 9 shows that the DPGMM-RNN hybrid model (RNNn) generates fewer cluster types than the DPGMM algorithm (DPGMM) and can compress the DPGMM clusters by ignoring the unstable ones with low probabilities, which makes the number of generated clusters nearer to the numbers of phonemes of the normal human languages.

5) Performance per Utterance: Besides the above comparisons between the v_measures of the whole corpus, we also did paired t-tests on the v_measures of the utterances of the timit test set. Except for the DPGMM-RNN hybrid model with 0 contextual frames (RNN0), all the other hybrid models with

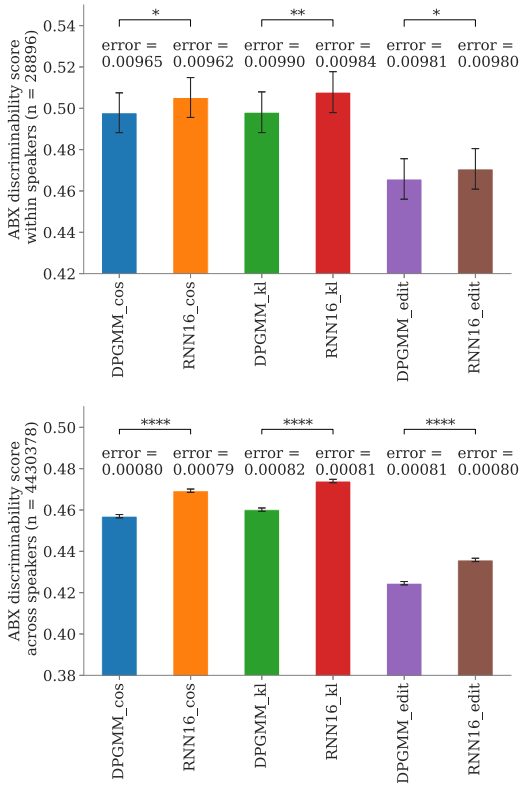


Fig. 11. Average ABX discriminability score within speakers (upper subfigure) and across speakers (lower subfigure) on TIMIT test corpus. We compared average ABX discriminability scores of all n triplets (A, B and X) between DPGMM algorithm (DPGMM) and DPGMM-RNN hybrid model of 16 contextual frames (RNN16) with cosine distance (cos), Kullback-Leibler divergence (kl), and Levenshtein distance (edit). Significance of paired t-test is indicated by stars: **** means $p \leq 0.0001$, ** means $p \leq 0.01$, and * means $p \leq 0.05$. Error bar is 95% confidence interval; error is annotated above.

longer contexts (RNN4, RNN8, and RNN16) significantly outperformed the DPGMM algorithm (DPGMM) on v_measures with p-value $p \leq 0.0001$.

D. Analysis of Cluster Discriminability of Phoneme Categories

As well as the information theory inspired by measures based on the relative frequency at the global corpus level, we measured the ability of our generated clusters for discriminating the triphone categories by computing the ABX discriminability scores [6] at the local segmental level.

Fig. 11 shows that the clusters from our proposed DPGMM-RNN hybrid model more effectively discriminate the phonemes than those from the DPGMM algorithm in ABX discriminability scores with three distances across and within speakers. The performance improvement shows statistical significance with the paired t-test. The error bar of the 95% confidence interval shows that the hybrid model achieved fewer errors than the DPGMM algorithm in ABX discriminability scores.

E. DPGMM-RNN Hybrid Model in Zerospeech 2019

Fig. 12 shows that the DPGMM-RNN hybrid model is better at discriminating phonemes (which is the decrease of the ABX error rate across different distances) and compressing representation (which is the decrease of the bit rate of the one-hot representation of clusters) compared to the DPGMM algorithm.

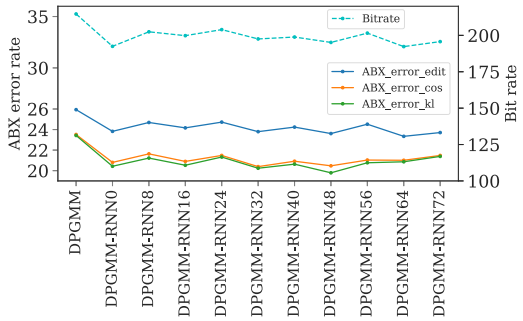


Fig. 12. ABX error rate and bit rate on English dataset of Zerospeech 2019 [27] (decreases simultaneously with stronger context modeling). Solid lines show ABX error rates with cosine distance, KL divergence and edit distance using the primary vertical axis; dashed line shows bit rate of generated clusters using secondary vertical axis. DPGMM means DPGMM clustering algorithm (without RNN contextual modeling); DPGMM-RNNn (RNNn) means hybrid model (with n RNN contextual frames).

As the number of contextual frames increases, the ABX error rates and the bit rates gradually decrease. We choose RNN48 (about three or four syllables [55] as RNN context) as the result of our DPGMM-RNN model for Zerospeech 2019 because the error rates with the cosine distance and KL divergence start increasing and that with the edit distance is still decreasing.

The Fig. 12 shows a sharp decrease in the bit rate and the ABX error rates between the DPGMM clustering algorithm and the DPGMM-RNN hybrid model. However, within the DPGMM-RNN models, increasing the length of the contexts slightly decreases the ABX error rate or the bit rate in Zerospeech 2019. The reason might be explained by its English training set, which only contains very short utterances, where the mean duration per utterance is 2.063 seconds, and the three longest utterance durations are 14, 7.99, and 7.82 seconds. When we increase the length of the context of the DPGMM-RNN hybrid model, we expect to capture both the acoustic structure of each phoneme and the statistical structure of a short sequence of several phonemes. This effect of modeling long contexts is relatively weak because most of the utterances of the English training corpus of Zerospeech 2019 are triphones instead of complete sentences of natural utterances. Longer contextual modeling doesn't show its full power on the dataset when all of the utterances are short.

The VQ-VAE [11], [56] and Factorized Hierarchical Variational Auto-encoder (FHVAE) [57] systems got the top ABX error rate results in Zerospeech 2019 [27]. We compared the system of the DPGMM-RNN hybrid model with these top systems with official toolkits from Zerospeech 2019 (Table I).

The best Zerospeech 2019 system used VQ-VAE [56] to quantize the MFCC acoustic features with several centroids. The system also uses a speaker-adversarial approach [40] to make the final representation speaker independent. Although our system of the DPGMM-RNN hybrid model (RNN48) got a relatively low ABX error rate, it had a slightly higher bit rate than the VQ-VAE based system.

TABLE I
ABX ERROR RATE AND BIT RATE OF DPGMM-RNN HYBRID MODELS (RNN48 AND BIRNN16) AND TOP MODELS FROM ZEROSPEECH 2019. THE PROVIDED ZEROSPEECH BASELINE USES DPGMM CLUSTERS TRAINED BY VARIATIONAL INFERENCE. VQ-VAE EXTRACTS DISCRETE REPRESENTATION WITH SPEAKER-ADVERSARIAL ENHANCEMENT (VQ-VAE). ADVERSARIAL MULTI-TASK LEARNING IS USED ON DPGMM CLUSTERS OBTAINED FROM ACOUSTIC FEATURES AFTER FHVAE TRANSFORMATION (FHVAE). OUR MODELS FIRST GET DPGMM CLUSTERS (DPGMM) FROM WHICH WE TRAIN THE DPGMM-RNN HYBRID MODEL USING THE UNIDIRECTIONAL RNN WITH 48 CONTEXTUAL FRAMES (RNN48). CONTEXTUAL MODELING OF THE HYBRID MODEL IS FURTHER ENHANCED USING THE BIDIRECTIONAL RNN WITH 16 CONTEXTUAL FRAMES (BIRNN16). NUMBERS OF CONTEXTUAL FRAMES OF DIFFERENT HYBRID MODELS ARE CHOSEN BASED ON THEIR LOWEST ABX ERROR RATES AND LOWEST BIT RATES ON THE ZEROSPEECH DATASET

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

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

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

The second best system first used the FHVAE extracted features to get DPGMM clusters. Those clusters and speaker ids are trained with adversarial multi-task learning to get a final representation. The system has its primary representation (FHVAE(b)) and an alternative (FHVAE(a)) [58].

The system FHVAE(a) gets a very low ABX error rate using continuous representation and high sampling rate to get more acoustic details, which also increases the error rate with edit distance and a very high bit rate. To decrease the bit rate and get discrete representation, the softmax outputs is converted to one-hot representations (FHVAE (b)). The system of our proposed DPGMM-RNN hybrid model got a lower ABX error rate across three provided distances and a lower bit rate than the FHVAE (b) system.

The official baseline [13] of Zerospeech 2019 uses compact representation (low bit rate) but sacrifices the discriminability of phonemes (relatively high ABX error rate).

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

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

F. Compression by Functional Load

Based on the nature of DPGMM clustering, the number of clusters keeps increasing as it sees more data. Although it is possible to limit the maximum number of DPGMM clusters by K by truncating the weights, picking a proper K requires additional knowledge of true data distribution, without which a non-convergence problem of Gibbs sampling might occur. Because of the acoustic complexity of speech, the DPGMM algorithm also tends to generate more clusters than phonemes in conventional human language [17].

We decreased the number of clusters of the DPGMM-RNN hybrid model by merging the units that can be easily disambiguated by their surroundings (the units with the low functional load; see Appendix for the algorithm). Fig. 13(a) shows that the ABX error rate changes very little after greedily merging 20 pairs of clusters with the lowest functional load.

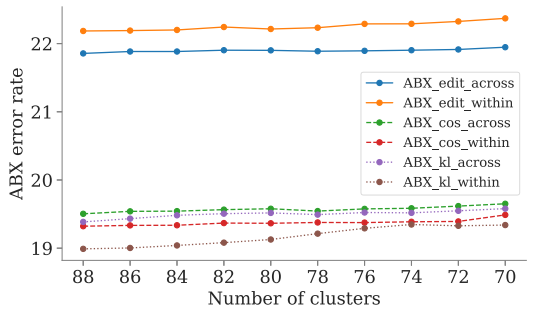
However, the functional load always exceeds zero, and the ABX error rate never decreases when we compressed the clusters with minial functional load. The TIMIT corpus, designed to be phonetically balanced and contextually complete, which makes contextual overlap for discovered segments and makes functional load never be zero. So we use the Xitsonga corpus with the same greedy strategy of merging pairs with the minimal functional load on DPGMM clusters [17]. The inner figure of Fig. 13(b) shows that the first 17 pairs of clusters bear the zero functional load, and we can compress the units and decrease the ABX error rate. Fig. 13(b) also shows that we can decrease the clusters from 188 to 58 without an obvious increase in the ABX error rate.

Fig. 13(c) shows the behavior of the language entropy and the functional load of the merged pairs when we used the minimal functional load principle for compression. As we merge more and more pairs with a larger functional load, more and more damage will be caused to the language's entropy.

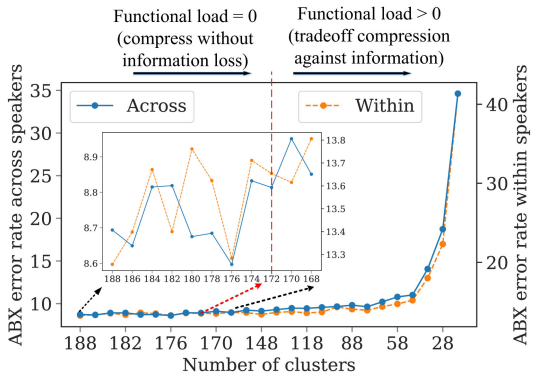
VI. DISCUSSION

We first question whether the fragmentation problem comes from the clustering difficulty of complex acoustical events because people use highly variated gestures to pronounce sounds with various manners, abstracted as distinctive features.

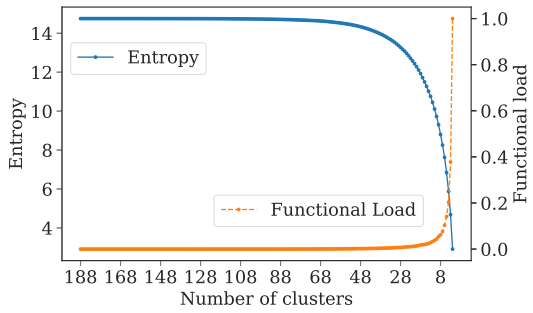
By exploring the fragmental level of different distinctive features by conditional perplexity (Fig. 7), we found that the DPGMM algorithm is worse at categorizing vowels than consonants because it generates more fragmental frames inside the phonemes. A similar situation happens in perception experiments where stably observing the construction of vowel categorization is more difficult than consonant categorization.



(a) ABX error rate after FL merges (TIMIT, RNN16)



(b) ABX error rate after FL merges (Xitsonga)



(c) Entropy and functional load after FL merges (Xitsonga)

Fig. 13. Subfigure (a): ABX error rate (across and within speakers with cosine, KL, and edit distances) after greedily applying minimum functional load merge on clusters from DPGMM-RNN hybrid model with 16 contextual frames on TIMIT corpus (RNN16). Subfigure (b): ABX error rate when we merge DPGMM cluster pairs with minimum functional load greedily on Zerospeech 2015 Xitsonga corpus. Red dashed line divides inner figure into the first 17 merged cluster pairs with zero functional load (left part) and the later pairs with non-zero functional load (right part). Subfigure (c): entropy of language and functional load of merged pairs when applying functional load compression on Xitsonga corpus.

For example, our perception jumps from one category to another when listening to the stimuli from /p/ to /t/ with equal acoustic changes of the consonants. However, our perception seems more continuous (hearing intra-phonemic variations) than categorical when listening to the stimuli from /i/, /I/ to /E/ with equal acoustic changes of the vowels [22].

The sensitivity of complex acoustic events causes the DPGMM algorithm to suffer from the fragmentation problem. Fig. 7 also shows that the fragments of the vowels are associated with the shapes and dynamics of formants. The fragments of the fricatives are associated with the energy concentrated at high frequencies, similar to noise.

Humans can perceive these noisy phonemes with the knowledge of language structure, even when we replace them with actual white noise [24]. This idea inspired us to propose such top-down contextual enhanced methods as RNN and functional load to capture the statistical structure of the acoustical segments for unsupervised phoneme discovery.

Our first proposal, the DPGMM-RNN hybrid model, explores how we can use high-level contextual information to relieve the problem of fragmentation. The hybrid model decreases the fragmental level (completeness increase) more than just using DPGMM clustering (Fig. 5, Fig. 8). Since we experimented on a longer context by taking more frames as RNN input, the fragmental level decreased more. With the same length of contextual frames, considering both the past and future context performances better than just considering one direction (Fig. 10).

Enhancing the contextual modeling by RNN helps remove the short-time fragmental segments without generating super-segments that cover several phonemes (Fig. 8), as shown by the decrease of the homogeneity and v-measure of our DPGMM-RNN model (Fig. 4, Fig. 8). The DPGMM-RNN hybrid model also compressed the segment systems by decreasing the number of clusters (Fig. 9).

The DPGMM-RNN hybrid model not only relieves the fragmentation problem but it also finds clusters that more accurately discriminate between phoneme categories. The hybrid model makes less ABX discrimination error (higher discriminability score) and performs more stable (tighter error bar) (Fig. 11). The DPGMM-RNN model also got a competitive performance in Zerospeech 2019 in discriminating English triphone segments (Table I).

Our second proposal, compression with functional load, explores how we can avoid excessive concentration on acoustic details while preserving the key information critical for communication. Our idea is to merge clusters that can be disambiguated by their surroundings. Merging such segments in complementary distribution with zero functional load typically causes no damage to conveying language's information. We found that we could discriminate phonemes better while compressing the segments with zero functional load, we also merged more segments with small functional load at the expense of a tiny decrease of ABX discriminability (Fig. 13).

The unsupervised phoneme discovery aims to find a phoneme sequence from each utterance. Perhaps the most related task is unsupervised phoneme segmentation [59], [60]. The two tasks differ in that unsupervised phoneme discovery only needs the correct phoneme order; segmentation emphasizes such detailed temporal information as the start and end times of each phoneme. Another related task is unsupervised speaker diarization [61]. In problem abstraction, both tasks map the acoustic signal to discrete categories, but unsupervised phoneme discovery ignores the speaker information and unsupervised speaker diarization ignores the segment information. Except for this difference, we believe both tasks can still learn from each other in their methods and evaluation metrics.

VII. Conclusion

To mimic the human perception bias of phonemes over acoustic signals, we proposed the DPGMM-RNN hybrid model to improve phoneme categorization, and proposed to merge unambiguous segments with low functional load. Results show that with the DPGMM-RNN hybrid model, we can relieve the fragmental problem and improve phoneme discriminability; with minimal merging of the functional load, we compressed the segment system, preserved the information, and retained the phoneme discriminability.

Appendix A

Computation of Functional Load

Assume a toy language is presented as a discrete label sequence 'A B C 12 13 14 D 12 13 14 E 12.' We want to include some context when computing the language's entropy, so instead of counting each single label, we count the string of labels with length 3, as we did in our experiment ('A B C' once, 'B C 12' once, ..., '12 13 14' twice, ..., '14 E 12' once). With the counts (1 1 1 2 1 1 1 1 1), we can compute the entropy (3.12) of the language by relative frequency (1/10, 1/10, 1/10, 2/10, 1/10, 1/10, 1/10, 1/10, 1/10).

1) Functional Load and Complementary Distribution: If we want to compute the functional load of units 'B' and '13,' which can be disambiguated by their environment (A, C) and (12, 14) because 'B' and '13' are in complementary distribution, we merge labels 'B' and '13' so the new language becomes 'A 13 C 12 13 14 D 12 13 14 E 12.' Since they have a complementary distribution, the counts of token string (1 1 1 2 1 1 1 1 1) and the entropy (3.12) of the new language do not change. Thus the functional load, as the loss the entropy, is ZERO. ((3.12 - 3.12) / 3.12 = 0).

2) Functional Load and Minimal Pair: If we want to compute the functional load of units 'E' and 'D,' which are in the same environment (14, 12), where 'E' and 'D' is a minimal pair. If we merge labels 'E' and 'D,' then the new language becomes 'A B C 12 13 14 D 12 13 14 D 12.' Because they are minimal pairs, the counts of token string (1 1 1 2 2 2 1) and the entropy (2.72) of the new language changes. Thus the functional load, as the loss of the entropy, is 0.13. ((3.12 - 2.72) / 3.12 = 0.13).

In our proposal, we merge these labels with minimal functional load to compress the clusters of a language. It is equivalent to merging the labels that are disambiguated by their surrounding environment [45]. Intuitively, merging labels in complementary distribution changes the tokens themselves without changing the distribution to compute the entropy of the language. Labels can be easily disambiguated from the surrounding labels bearing zero or very small functional load.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number JP17H06101.

References

[1] C.-y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. 50th Annu. Meet. Assoc. Comput. Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[2] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proc. 46th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.: Short Papers*. Association for Computational Linguistics, 2008, pp. 165–168.

[3] M. Huijbregts, M. McLaren, and D. Van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 4436–4439.

[4] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011, pp. 401–406.

[5] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 186–197, Jan. 2008.

[6] M. Versteegh *et al.*, "The zero resource speech challenge 2015," in *Proc. INTERSPEECH*, 2015, pp. 3169–3173.

[7] 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.

[8] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An Auto-encoder based approach to unsupervised learning of subword units," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 7634–7638.

[9] L. Badino, A. Mereta, and L. Rosasco, "Discovering discrete subword units with binarized Autoencoders and hidden-Markov-model encoders," in *Proc. INTERSPEECH*, 2015, pp. 3174–3178.

[10] 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.

[11] 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 2019," 2019, *arXiv:1905.11449*.

[12] 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.* Springer, 2017, pp. 169–180.

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

[14] 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.

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

[16] 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.

[17] B. Wu, S. Sakti, J. Zhang, and S. Nakamura, "Optimizing DPGMM clustering in zero resource setting based on functional load." in *Proc. Spoken Lang. Technol. Under-Resourced Lang.*, 2018, pp. 1–5.

[18] 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/Recon Techn. Rep. n*, vol. 93, 1993.

[19] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, no. 2, pp. 175–184, 1952.

[20] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[21] C. F. Hockett, *A Manual of Phonology*. Waverly Press, 1955, no. 11.

[22] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code." *Psychol. Rev.*, vol. 74, no. 6, p. 431, 1967.

[23] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.

[24] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, no. 3917, pp. 392–393, 1970.

[25] W. F. Ganong, "Phonetic categorization in auditory word perception.," *J. Exp. Psychol.: Hum. Percep. Perform.*, vol. 6, no. 1, p. 110, 1980.

[26] M. A. Pitt and J. M. McQueen, "Is compensation for coarticulation mediated by the lexicon?," *J. Memory Lang.*, vol. 39, no. 3, pp. 347–370, 1998.
[27] E. Dunbar *et al.*, "The zero resource speech challenge 2019: TTS without T," 2019, *arXiv:1904.11469*.
[28] J. Chang and J. W. Fisher III, "Parallel sampling of DP mixture models using sub-cluster splits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 620–628.
[29] 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.
[30] Y. W. Teh, "Dirichlet process," *Encyclopedia Mach. Learn.*, pp. 280–287, 2010.
[31] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 4, pp. 769–773, 1955.
[32] A. Fourtassi and E. Dupoux, "A rudimentary lexicon and semantics help bootstrap phoneme acquisition," in *Proc. 18th Conf. Comput. Natural Lang. Learn.*, 2014, pp. 191–200.
[33] A. Fourtassi, E. Dunbar, and E. Dupoux, "Self-consistency as an inductive bias in early language acquisition," in *Proc. Annu. Meet. Cognit. Sci. Soc.*, vol. 36, no. 36, 2014.
[34] N. Feldman, T. Griffiths, and J. Morgan, "Learning phonetic categories by learning a lexicon," in *Proc. Annu. Meet. Cognit. Sci. Soc.*, vol. 31, no. 31, 2009.
[35] M. Johnson, T. L. Griffiths, and S. Goldwater, "Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 641–648.
[36] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1385–1392.
[37] M. Heck, S. Sakti, and S. Nakamura, "Iterative training of a DPGMM-HMM acoustic unit recognizer in a zero resource scenario," in *Spoken Lang. Technol. Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 57–63.
[38] G. A. Miller and W. G. Taylor, "The perception of repeated bursts of noise," *J. Acoust. Soc. Amer.*, vol. 20, no. 2, pp. 171–182, 1948.
[39] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychology Rev.*, vol. 63, no. 2, p. 81, 1956.
[40] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," 2018, *arXiv:1804.02812*.
[41] T. Tsuchiya, N. Tawara, T. Ogawa, and T. Kobayashi, "Speaker invariant feature extraction for zero-resource languages with adversarial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2381–2385.
[42] M. Heck, S. Sakti, and S. Nakamura, "Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering," in *Proc. INTERSPEECH*, 2016, pp. 1310–1314.
[43] 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.
[44] J.-S. Zhang, X.-H. Hu, and S. Nakamura, "Using mutual information criterion to design an efficient phoneme set for Chinese speech recognition," *IEICE TRANSACTIONS Inf. Syst.*, vol. 91, no. 3, pp. 508–513, 2008.
[45] W.-Y. Wang, "The measurement of functional load," *Phonetica*, vol. 16, no. 1, pp. 36–54, 1967.
[46] J. Zhang, W. Li, Y. Hou, W. Cao, and Z. Xiong, "A study on functional loads of phonetic contrasts under context based on mutual information of chinese text and phonemes," in *Proc. 7th Int. Symp.Chin. Spoken Lang. Process.*, 2010, pp. 194–198.
[47] D. Surendran and P. Niyogi, "Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals," *Amsterdam studies in the theory and history of linguistic science series 4*, vol. 279, p. 43, 2006.
[48] M. André, "Economie des changements phonétiques," *Berne: Francke*, 1955.
[49] 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. 1–5.
[50] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica sinica*, pp. 639–650, 1994.
[51] C. F. Hockett, "The quantification of functional load," *Word*, vol. 23, no. 1-3, pp. 300–320, 1967.
[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] T. Schatz, "ABX-discriminability measures and applications," Ph.D. dissertation, 2016.
[54] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.
[55] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Temporal properties of spontaneous speech-a syllable-centric perspective," *J. Phonetics*, vol. 31, no. 3-4, pp. 465–485, 2003.
[56] A. van den Oord *et al.*, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6306–6315.
[57] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1878–1889.
[58] S. Feng, T. Lee, and Z. Peng, "Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling," 2019, *arXiv:1906.07234*.
[59] A. H. H. N. Torbati, J. Picone, and M. Sobel, "Speech acoustic unit segmentation using hierarchical Dirichlet processes," in *Proc. INTERSPEECH*, 2013, pp. 637–641.
[60] P. Baljekar, S. Sitaram, P. K. Muthukumar, and A. W. Black, "Using articulatory features and inferred phonological segments in zero resource speech processing," in *Proc. INTERSPEECH*, 2015, pp. 3194–3198.
[61] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2015–2028, Oct. 2013.

Bin Wu received the B.E. degree in computer science from the Jiangsu University of Science and Technology and the M.E. degree in computer applied technology from the Beijing Language and Culture University. He is the Ph.D. Student with the Nara Institute of Science and Technology, Ikoma, Japan. His research interests include the 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. In 2000, she was the recipient of the DAAD-Siemens Program Asia 21st Century Award to study communication technology at the University of Ulm, Ulm, Germany, where she received the M.Sc. degree in 2002. During her thesis work, she worked with the Speech Understanding Department, the Daimler-Chrysler Research Center, Ulm. Between 2003 and 2009, she worked as a Researcher with the ATR SLC Labs, Kyoto, Japan, and from 2006 to 2011, as an Expert Researcher with the NICT SLC Groups, Japan. While working with ATR and NICT, she continued her study from 2005 to 2008 with the Dialog Systems Group, University of Ulm and received the Ph.D. degree in 2008. She has been actively involved in such collaboration activities as the Asian Pacific Telecommunity Project from 2003 to 2007 and ASTAR and USTAR from 2006 to 2011. From 2009 to 2011, she was a Visiting Professor with the Computer Science Department, University of Indonesia, Depok, Indonesia. From 2011, she has been an Assistant Professor with the Augmented Human Communication Laboratory, NAIST, Japan. She was also a Visiting Scientific Researcher of INRIA Paris Rocquencourt, France from 2015 to 2016, under the 'JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation.' Since January 2018, she serves as a Research Associate Professor at NAIST and a Research Scientist at RIKEN AIP, Japan. She is a Member of JNS, SFN, ASJ, ISCA, IEICE, and IEEE. Furthermore, she is currently a Committee Member of IEEE SLTC (2021-2023) and an Associate Editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing (2020-2023). She is also the officer of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a Board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU). 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.

Jinsong Zhang (Member, IEEE) received the B.E. degree in electronic engineering from the Hefei University of Technology, China in 1989, the M.E. degree in electronic circuit, signal and system from the University of Science and Technology of China (USTC) in 1992, and the Ph.D. degree in information and communication engineering from the University of Tokyo, Japan in 2000. From 1992 to 1996, he worked as 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, etc.

Satoshi Nakamura (Fellow, IEEE) received the B.S. degree from the Kyoto Institute of Technology in 1981 and the Ph.D. degree from Kyoto University in 1992. He is a Professor with the Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, Project Leader with the Tourism Information Analytics Team of RIKEN, Center for Advanced Intelligence Project AIP, the Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He was an Associate Professor with the Graduate School of Information Science, the Nara Institute of Science and Technology from 1994–2000 and the Director with the ATR Spoken Language Communication Research Laboratories in 2000–2008 and Vice President with ATR in 2007– 2008. He was the Director General with the Keihanna Research Laboratories and the executive director of the Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009–2010. He is currently the Director with the Augmented Human Communication laboratory and a Full Professor with 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 Yamashita Research Award, the Kiyasu Award from the Information Processing Society of Japan, the Telecom System Award, the AAMT Nagao Award, the Docomo Mobile Science Award in 2007, and the ASJ Award for Distinguished Achievements in Acoustics. He received the 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 also received the LREC Antonio Zampolli Award in 2012. He has been an Elected Board Member of the International Speech Communication Association, ISCA, since June 2011, an *IEEE Signal Processing Magazine* Editorial Board Member since April 2012, an IEEE SPS Speech and Language Technical Committee Member since 2013.