PAPER

Learning Supervised Feature Transformations on Zero Resources for Improved Acoustic Unit Discovery

Michael HECK[†a)], ***Nonmember***, **Sakriani SAKTI**[†], ***and*** **Satoshi NAKAMURA**[†], ***Members***

SUMMARY In this work we utilize feature transformations that are common in supervised learning without having prior supervision, with the goal to improve Dirichlet process Gaussian mixture model (DPGMM) based acoustic unit discovery. The motivation of using such transformations is to create feature vectors that are more suitable for clustering. The need of labels for these methods makes it difficult to use them in a zero resource setting. To overcome this issue we utilize a first iteration of DPGMM clustering to generate frame based class labels for the target data. The labels serve as basis for learning linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) and feature-space maximum likelihood linear regression (fMLLR) based feature transformations. The novelty of our approach is the way how we use a traditional acoustic model training pipeline for supervised learning to estimate feature transformations in a zero resource scenario. We show that the learned transformations greatly support the DPGMM sampler in finding better clusters, according to the performance of the DPGMM posteriorgrams on the ABX sound class discriminability task. We also introduce a method for combining posteriorgram outputs of multiple clusterings and demonstrate that such combinations can further improve sound class discriminability.

key words: *acoustic unit discovery, Bayesian nonparametrics, feature transformation, unsupervised subword modeling, zero resource*

1. Introduction

We speak of a *zero resource scenario* in the speech processing domain, when large amounts of labeled training data, parallel data, and expert knowledge about the target language are unavailable for techniques of supervised learning. State-of-the-art machine learning methods that typically rely on accurate, hand-crafted labels for training can not be applied easily in such a scenario. Unsupervised methods try to circumvent the need for precise labels and extensive expert knowledge, but despite significant advances in this field, it still remains a challenge to imitate human capacities of learning models of spoken language.

Phonologists approach a new and unseen language by defining a set of acoustic units to fully cover the underlying sound repertoire. Machine learning approaches to this are pattern matching on raw audio data [1], [2] and unsupervised learning of models [3]. These techniques have been successfully applied to solve tasks such as spoken term detection [4], topic segmentation [5] or document classification [6].

Manuscript received May 29, 2017.
Manuscript revised September 13, 2017.
Manuscript publicized October 20, 2017.

[†]The authors are with the Augmented Human Communication Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
a) E-mail: michael-h@is.naist.jp

DOI: 10.1587/transinf.2017EDP7175

Recently, evaluations such as the zero resource speech challenge [7] specialize in learning a new language from scratch and without any prior supervision. The challenge defines the above-mentioned task as *unsupervised subword modeling*, where the objective is to construct a representation of speech sounds that is robust to variation within and across speakers and that maximizes class discrimination [7]. This task was tackled by a spectrum of contributions: [8] applies a correspondence auto-encoder to learn efficient representations with the help of matched word pairs generated by an unsupervised term discovery (UTD) system. [9] makes use of a deep auto-encoder that applies a threshold at the encoding layer to generate a binary representation of speech frames. [10] proposes a Siamese DNN training framework that takes the frames of UTD word pairs as input and minimizes the distance between frames of the same class and maximizes it between frames of different classes.

Model complexity usually is not known a priori when dealing with new data sets and where estimation is not possible due to the lack of development data. Bayesian nonparametric models can be a good choice in such cases, as they automatically adjust the model complexity given some data. Bayesian models have already been successfully applied to other speech processing tasks such as unsupervised lexical clustering [11]. Chen et al. [12] take a Bayesian nonparametric approach to unsupervised subword modeling and cluster MFCCs with their derivatives by inferring a Dirichlet process Gaussian mixture model (DPGMM). A major drawback of clustering MFCC feature vectors is that they are not implicitly designed to maximize discriminability.

On the other hand, traditional supervisedly trained speech processing systems typically utilize a whole array of feature transformations to optimize the feature vectors towards discriminability. Linear discriminant analysis (LDA) [13] for instance is a standard technique to minimize intra-class discriminability, to maximize inter-class discriminability and to extract relevant informations from high-dimensional features spanning larger contexts. Maximum likelihood linear transforms (MLLT) [14], [15] is a common method to de-correlate feature components, and feature-space maximum likelihood linear regression (fMLLR) [16], [17] is widely used for speaker adaptation. Class discriminating properties are critical for clustering methods like the one used in [12], and adaptive feature transformations can help reduce variability. However, methods such as LDA require class labels for estimating the transformations, which makes it difficult to use them in zero resource sce-

narios, where the class identities and even their amount are unknown.

In this work, we present a new approach to optimizing the input of a particular Bayesian non-parametric clustering method, the DPGMM sampler. The novelty of this work is the way how we use traditional methods of supervised learning for unsupervised feature transformation estimation in a zero resource scenario, i.e., without having any prior labels. We demonstrate that we can learn various feature transformations on automatically generated labels instead, and that these transformations produce feature vectors that considerably improve clustering performance. There has been work that utilize k-means clustering to automatically obtain pseudo labels for LDA estimation [18], [19]. However, unlike in these studies we overcome the limitation of having to predefine the size of the label set by using the non-parametric DPGMM sampler itself to generate initial labels for our untranscribed data. These labels serve as basis for learning LDA, MLLT and fMLLR transformations in an entirely unsupervised fashion. For that, we efficiently utilize a classic acoustic model training pipeline, which makes our approach easy-to-use. The original input is transformed and clustered again with the DPGMM sampler. Our experiments show that the feature transformations greatly help improve cluster quality and that using multiple transformations produces the best results. We demonstrate the effectiveness of our method on two very different data sets that vary in size, language and speaking style. Additionally, we introduce a method for combining the results of multiple DPGMM samplings on posteriorgram level and show that such combination can boost sound class discriminability even further. We believe our transformation based approach to optimizing feature vectors for clustering is universal and will be useful for other zero and low resource tasks as well.

2. Dirichlet Process Gaussian Mixture Model

DPGMMs (also known as infinite GMMs) extend finite mixture models by the aspect of automatic model selection: The model finds its complexity automatically given the data. Model inference is typically sample based using a Markov chain Monte Carlo (MCMC) scheme such as Gibbs sampling. The sampler used here combines a restricted Gibbs sampler with a split/merge sampler. For more in-depth informations, please refer to [12] and [20].

2.1 Generative Process

Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set of observations. The generative process of X given a DPGMM is as follows:

- Mixing weights $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_k\}$ are generated according to a stick-breaking process
- GMM parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$ are generated according to an Normal-inverse-Wishart (NIW) distribution as $\mathrm{NIW}(m_k, S_k, \kappa_k, \nu_k)$ as prior distribution
- A label z_i is assigned to every data point $\boldsymbol{x}_i$, according to the mixing weights $\boldsymbol{\pi}$
- A data point $\boldsymbol{x}_i$ is generated according to the z_i-th GMM component

$\theta_k = \{\mu_k, \Sigma_k\}$ are Gaussian parameters, and the parameter set of the prior Normal-inverse-Wishart (NIW) distribution consists of a prior m_0 for μ_k, a prior S_0 for Σ_k, the belief-strength κ_0 in m_0 and the belief-strength ν_0 in S_0.

2.2 Inference

The parallelizable sampler alternates between a non-ergodic restricted Gibbs sampler and a split/merge sampler to form an ergodic MCMC sampler.

Restricted Gibbs sampling allows labels z_i to be sampled from a finite set of labels Z. By definition of the DPGMM, the distribution of the mixture weights follows a Dirichlet distribution.

Split/merge sampling performs operations on the existing components. To provide good split candidates, each component is augmented with two sub-clusters with mixing weights $\pi_{k,l}, \pi_{k,r}$ and parameter sets $\theta_{k,l}, \theta_{k,r}$, and each observation of a component is augmented with a sub-cluster label $z_{sub_i} \in l, r$.

The Split/merge sampler proposes split and merge moves in a Metropolis-Hastings fashion. A Hastings ratio H is computed according to the momentary assignment of observations of a component to its sub-clusters, and a move is accepted with a probability $\min(1, H)$. For the merge step, merges of randomly picked components are proposed.

3. Multi-Stage Clustering for Acoustic Unit Discovery

Our approach to improving the quality of DPGMM based speech feature vector clustering is realized by a multi-stage framework. This framework utilizes multiple feature transformations in conjunction to benefit from additive effects. We want to make use of speech feature transformations which are well-established for rich-resource languages to optimize the input features towards discriminability. A wide range of transformations can be applied to features for this purpose, with favorable effects such as dimensional reduction, feature de-correlation or adaptation to certain conditions in order to minimize variability. Because we are situated in a zero-resource scenario, we exploit these transformations in an unsupervised fashion.

We utilize a standard pipeline for supervised acoustic model training, where feature transformations are conveniently estimated during the course of the training process to obtain the transformations. The advantage of this is that well-established pipelines already exist and are ready to use. The disadvantage is the requirement of labels for training.

To overcome the issue of not having labels in a zero-resource scenario, we propose using a multi-stage strategy that alternates between feature vector clustering, label generation and transformation estimation via model training. We use the DPGMM sampler to generate initial labels for

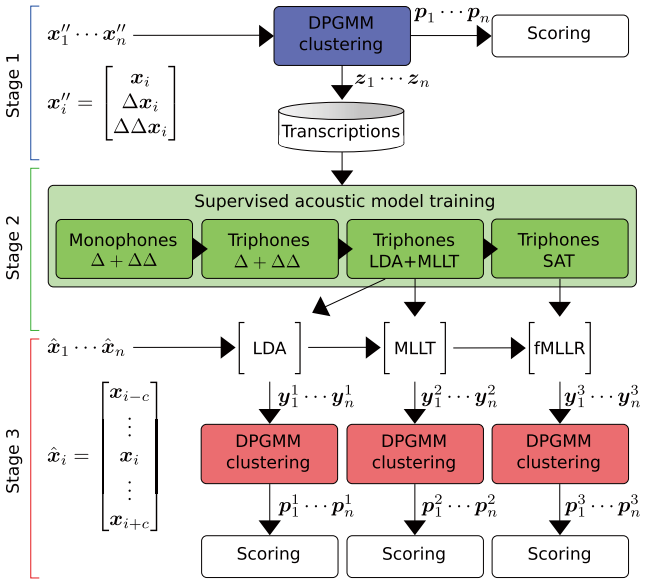

Fig. 1 Scheme of the multi-stage clustering for acoustic unit discovery. In stage 1, standard features are clustered. From frame based class labels, utterance based transcriptions are generated. In stage 2, feature transformations are estimated with the help of an acoustic model training pipeline and the automatic transcriptions. In stage 3, features are transformed with one or more transformations before clustering them by sampling a DPGMM.

our untranscribed data by clustering standard feature vectors. These automatic labels serve as basis for feature transformation estimation by unsupervised acoustic model training. The transformations are then applied to the standard feature vectors prior to a second run of DPGMM based clustering. We explain the individual stages of our framework in detail in the following Subsections. A graphical overview is given in Fig. 1.

3.1 Stage 1: Clustering Standard Feature Vectors

The DPGMM as a Bayesian non-parametric model has the convenient property to automatically find an optimal number of classes given a set of data during sampling. We use this property and run an initial clustering on standard feature vectors with derivatives ($\boldsymbol{x}_i''$) to get a set of class labels and the hypothesized class membership z_i for all n speech frames. These classes are simply named with the numeric ID of the Gaussian distribution that most likely produced the respective feature vector.

3.2 Stage 2: Transformation Estimation

The output of the previous step is frame-wise class labels for the data. We collapse the labels for each utterance by compressing all subsequent tokens of the same type to a single token, i.e., a sequence of labels 1-1-2-2-2-3-4-4-4 becomes 1-2-3-4. This is done to imitate transcriptions based on phone-like units. We use these transcriptions for transformation estimation by running an out-of-the-box acoustic model training pipeline. We use a 3-state HMM topology with a skip from the first state to the next HMM to guarantee that an alignment is always found, since we cannot guarantee that every label in the transcription covers at least 3 frames. The training is initialized with a flat-start, i.e., by context-independent monophone training starting with an equally spaced alignment. Then we subsequently train context dependent tri-phones, where during training we estimate transformations based on LDA, MLLT and fMLLR.

LDA is a well-known linear transformation that we use to minimize intra-class discriminability and maximize inter-class discriminability of speech features. Estimating the LDA transformation requires the feature vectors themselves and their respective class labels. With our pipeline, we create alignments between utterance HMMs and the automatic textual labels from the previous step and use the HMM states as classes for the LDA.

We compute the LDA for stacked feature vectors ($\hat{\boldsymbol{x}}_i$), where we use a context of c, meaning that we stack the c left and c right feature vectors on top of the current vector, which is the center vector. Context is an important source of information to correctly classify speech features. Feature stacking can cover a much larger context than appending the first and second derivatives, for instance. Dimensional reduction of these high-dimensional vectors is done by omitting lower-ranked coefficients after applying the transformation. Lower dimensional feature vectors encapsulate relevant information more efficiently and help keep the clustering feasible.

MLLT is computed for distributions of speech observations in the HMMs of speech recognizers. The main purpose of MLLT in speech recognition systems is to force the features into a space where diagonal modeling is suitable, which greatly reduces complexity and thus simplifies computing the model parameters [14]. The state-dependent transformations are estimated so that the likelihood of the adaptation data is maximized. Our motivation to use MLLT is to capture correlations between feature vector components.

fMLLR is an algorithm for speaker adaptive training (SAT). The idea of SAT is to capture inter-speaker variability in speaker dependent transformations and to generate speaker independent state distributions instead. Since the transformations are applied in the feature space, the resulting feature vectors are expected to show lower variability across speakers. The transformations are estimated based on alignments with speaker-independent features so that the likelihoods are maximized. We apply fMLLR in the zero resource scenario because we expect the transformations to help eliminate variance caused by multiple speakers, which should intuitively aid the clustering process.

3.3 Stage 3: Clustering Transformed Feature Vectors

We extract the transformations learned in the previous step as transformation matrices, which can be readily applied to the feature vectors $\hat{\boldsymbol{x}}_i$ prior to a second run of DPGMM based clustering. As illustrated in Fig. 1, we can extract new

feature vectors $\boldsymbol{y}_i$ by using one (LDA) up to three transformations (LDA+MLLT+fMLLR) in conjunction. After applying one or more transformations, we perform the frame based DPGMM clustering. We compare the clustering quality using the untransformed features with the clustering quality using each of the transformed features. For that we first extract m sets of GMM posteriorgrams $\boldsymbol{p}_i^m$ for our data given each of the m DPGMMs and then score each of these posteriorgram sets (see Sect. 5.2).

4. Posteriorgram Combination

In this Section, we describe the method that we developed to combine the output of multiple clusterings. System combination on hypothesis level is a popular method in speech processing to further improve the output quality. Inspired by the idea, we developed a method to combine the output of multiple clusterings on posteriorgram level. The method is formally expressed in Algorithm 1.

The output of each DPGMM m can be represented as a set of posteriorgrams $P_m = \{\boldsymbol{p}_1^m, \ldots, \boldsymbol{p}_n^m\}$ with one posteriorgram for each of the n speech frames (see Eq. (2)). Generally, combining multiple sets of posteriorgrams $\mathcal{P} = \{P_1, \cdots, P_m\}$ is straightforward. For each frame i, we add together the m individual posteriorgrams $\{\boldsymbol{p}_i^1, \ldots, \boldsymbol{p}_i^m\}$ (Operation 13) and normalize the new vectors (Operation 15):

$$\hat{\boldsymbol{p}}_i = \frac{1}{m} \sum_{k=1}^{m} \boldsymbol{p}_i^k \tag{1}$$

The result is a new set of posteriorgrams $\hat{P} = \{\hat{\boldsymbol{p}}_1, \ldots, \hat{\boldsymbol{p}}_n\}$.

However, for the non-parametric DPGMM, the amount of found classes and thus the dimensionality of posteriorgram vectors differs for each clustering run. Therefore, a mapping between any two sets of posteriorgrams is needed. Given m sets of posteriorgrams, we randomly pick one of these sets as target set $P_{\mathtt{tgt}}$ (Operation 1), and consider all other sets as source sets, each denoted as $P_{\mathtt{src}}$ (Operation 4).

The mapping from $P_{\mathtt{src}}$ to the space of $P_{\mathtt{tgt}}$ for any source/target pair works as follows: We first convert all frame-wise posteriorgrams in $P_{\mathtt{tgt}}$ into frame-wise labels $l_{\mathtt{tgt}}$ by taking the numeric ID of the class with the highest probability as label (Operation 2). Knowing the frame-wise labels, we can represent each utterance in our data as sequences of labels. We do the same given all the posteriorgrams in $P_{\mathtt{src}}$ (Operation 5). For each utterance we now have a pair of label sequences, which we align to count the label co-occurences $t_{\mathtt{cooc}}$ (Operation 6). Given the counts we identify the single most probable "translation" for each class ID, which we keep in a mapping table $t_{\mathtt{1best}}$ (Operation 7). With the mapping table it is possible to rearrange the posteriorgram vector elements for all $\boldsymbol{p} \in P_{\mathtt{src}}$ to match the posteriorgram vector layout of $P_{\mathtt{tgt}}$ (Operation 10). Note that there can be many-to-one mappings in case the posteriorgrams in $P_{\mathtt{src}}$ have higher dimensionality than the ones in $P_{\mathtt{tgt}}$. For an intuitive example of mapping a single posteriorgram, see Fig. 2.

Algorithm 1 Posteriorgram combination

Require: Set $\mathcal{P} = \{P_1, \cdots, P_m\}$ of sets of posteriorgrams
Ensure: Combined set of posteriorgrams $\hat{P}$
1: $P_{\mathtt{tgt}} \leftarrow$ random set from $\mathcal{P}$
2: $l_{\mathtt{tgt}} \leftarrow$ generate labels from posteriorgrams $P_{\mathtt{tgt}}$
3: $\hat{P} \leftarrow P_{\mathtt{tgt}}$
4: **for all** $P_{\mathtt{src}} \in \mathcal{P} \setminus P_{\mathtt{tgt}}$ **do**
5: $\quad l_{\mathtt{src}} \leftarrow$ generate labels from posteriorgrams $P_{\mathtt{src}}$
6: $\quad t_{\mathtt{cooc}} \leftarrow$ count label co-occurrences in align($l_{\mathtt{src}}, l_{\mathtt{tgt}}$)
7: $\quad t_{\mathtt{1best}} \leftarrow$ keep 1-best mapping from $t_{\mathtt{cooc}}$
8: $\quad \hat{P}_{\mathtt{src}} \leftarrow \{\}$
9: $\quad$ **for all** $\boldsymbol{p} \in P_{\mathtt{src}}$ **do**
10: $\quad\quad \boldsymbol{p}_{\mathtt{map}} \leftarrow$ map $\boldsymbol{p}$ to space of $P_{\mathtt{tgt}}$ using $t_{\mathtt{1best}}$
11: $\quad\quad$ add $\boldsymbol{p}_{\mathtt{map}}$ to set $\hat{P}_{\mathtt{src}}$
12: $\quad$ **end for**
13: $\quad \hat{P} \leftarrow$ add together pair-wise posteriorgrams in $\hat{P}_{\mathtt{src}}$ and $\hat{P}$
14: **end for**
15: $\hat{P} \leftarrow$ normalize $\hat{P}$

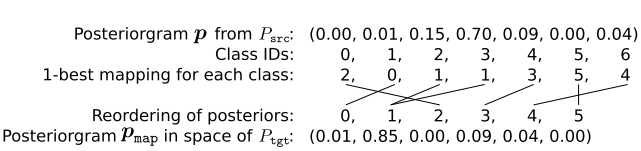

Fig. 2 Example of mapping one posteriorgram $\boldsymbol{p}$ from the source set $p_{\mathtt{src}}$ to the space of target set $P_{\mathtt{tgt}}$. The 1-best mappings in the mapping table $t_{\mathtt{1best}}$ are used to re-arrange the posteriorgram vector elements to match the posteriorgram vector layout of the target set $P_{\mathtt{tgt}}$. There can be many-to-one mappings in case the posteriorgrams in $P_{\mathtt{src}}$ have higher dimensionality than in $P_{\mathtt{tgt}}$.

5. Experimental Setup

5.1 Data

The database for all our experiments are the two official data sets of the Interspeech zero resource speech challenge 2015 [7], which greatly vary in size, language and speaking style. One set contains spontaneous, conversational interview-style American English (4h 59min), extracted from the Buckeye corpus [21]. The other set contains carefully uttered, read speech in Xitsonga (2h 29min), a southern African Bantu language. The latter is an excerpt of the NCHLT corpus [22]. All speech segments contain non-overlapping speech of exactly one speaker and are free of non-human noises and pauses. We extract about 1.7M frames for English and 0.8M frames for Xitsonga to cluster.

5.2 Evaluation Method

The evaluation metric we use to measure the cluster quality is based on the minimal pair ABX phone discriminability between phonemic minimal pairs [23]. We score GMM posteriorgrams that are computed for each speech frame after clustering, where the posterior probability of the cluster c_k, given an observation $\boldsymbol{x}_i$ is computed as

$$p(c_k|\boldsymbol{x}_i) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\theta_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\theta_j)} \tag{2}$$

where K is the total number of components in the DPGMM

and $\boldsymbol{p}_i = (p(c_1|\boldsymbol{x}_i), \ldots, p(c_K|\boldsymbol{x}_i))$ forms the posteriorgram for observation $\boldsymbol{x}_i$. θ are the Gaussian parameters, and π are the mixing weights (see Sect. 2).

Let A and B be speech representations of sound categories a and b, and X be of either a or b. The ABX phone discrimination error is

$$c(a,b) = 1 - \frac{1}{|a| \cdot |b| \cdot (|a|-1)} \sum_{A \in a} \sum_{B \in b} \sum_{X \in a \setminus \{A\}} \left(\delta_{d(A,X)<d(B,X)} + \frac{1}{2}\delta_{d(A,X)=d(B,X)}\right) \quad (3)$$

where $\delta_{(\cdot)}$ is an indicator function that equals to 1 if the condition $(\cdot)$ holds true and is 0 otherwise, and $d(\cdot,\cdot)$ is the dynamic time warping (DTW) distance defined over sequences of frame based feature vectors (in this case posteriorgrams). As in Schatz et al. [23], we use the Kullback-Leibler divergence to compute the DTW distances.

The idea of the ABX test is as follows. Given a phone based reference transcription of the test data, and the posteriorgrams coming from the DPGMM sampler, we can identify sequences of posteriorgrams that represent the same phone-triplets, between which we can compute distances. For example, if A and X are two different sequences of posteriorgrams that represent the triplet "b-a-g", and B is another sequence that represents "b-e-g", then the distance between A and X should be smaller than between B and X. If this is not the case, then this counts as a discrimination error. We collect the errors for all possible pairings of central phones. The errors are averaged over all contexts for a given pair of central phones and then over all pairs of central phones. Moreover, we compute the errors within speakers (i.e., the average phone discriminability error for each speaker specific portion of the test data) and across speakers.

The references for English and Xitsonga contain 165k and 72k phone-triplet annotations for 39 and 53 unique center phones, respectively. On average, there are 4.2k and 1.3k samples for each English and Xitsonga phone, respectively. This allows reliable discriminability error analyses.

5.3 Tools

We use the Kaldi speech recognition toolkit [24] to extract speech feature vectors for a frame length of 25 milliseconds and frame shift of 10 milliseconds. We apply mean variance normalization (MVN) and vocal tract length normalization (VTLN). The VTLN is done by learning a universal background model on the full data set for each language and subsequent training of a model for the extraction of warp factors. That gives us warp factors for the target data already, but the resulting models could be used for future unseen data to extract warp factors without the overhead of any re-training. All AMs used in our framework are likewise trained with Kaldi, following a standard scheme for speaker adaptive training (Kaldi recipe s5). Since we work in a zero resource scenario, all parameters that can be tuned are set to default values. We use the same parameters as Chen et al. [12] during the DPGMM sampling to ensure comparability. The sampling is done for 1500 iterations, and the priors are set so that m_0 is the global mean, S_0 is the global covariance, $\kappa_0 = 1$, and $\alpha = 1$. The value of ν_0 slightly varies and is set to the toolkit's default of $\nu_0 = D + 3$, where D is the dimension of the input feature vectors.

6. Evaluation

6.1 Baselines

The baseline discriminability error rates were produced by clustering 39 dimensional MFCC or PLP vectors with first and second order derivatives (MFCC+Δ+ΔΔ) with the DPGMM sampler and extracting the GMM posteriorgrams, which is the method of Chen et al. [12]†.

For comparison, we computed another set of baseline discriminability error rates by using principal component analysis (PCA) [25], [26] to transform the feature vectors prior to the DPGMM sampling. PCA is an entirely unsupervised method to de-correlate variables with an orthogonal linear transformation and is closely related to LDA, which makes it a fair basis of comparison for the effect of the supervised transformations that we learn without prior supervision. The baseline numbers are found in Table 1.

6.2 PCA vs. LDA

Figure 3 plots discriminability errors of GMM posteriorgrams that were extracted after clustering PCA or LDA transformed feature vectors. The graphs show the performance with regards to the output dimensionality of the transformations, i.e., how many coefficients are used after transforming the features vectors.

Table 1 Summary of the experimental results. The table contrasts Chen et al.'s best performance (row 1), our baseline performance (row 2) (for details about the differences see Sect. 6.1), results using various feature transformations, and the best posteriorgram-level combination (Comb. *V*, see Table 2) (row 12).

	English		Xitsonga	
Features	within	across	within	across
MFCC+Δ+ΔΔ ([12])	10.8	16.3	9.6	17.2
MFCC+Δ+ΔΔ	12.2	19.5	8.9	14.2
MFCC+PCA	11.7	19.2	9.8	16.4
MFCC+LDA	11.0	16.6	8.7	13.2
MFCC+LDA+MLLT	11.0	16.5	8.7	13.1
MFCC+LDA+MLLT+fMLLR	11.0	16.0	8.6	12.7
PLP+Δ+ΔΔ	11.8	19.6	8.5	13.9
PLP+PCA	11.7	18.4	8.7	14.6
PLP+LDA	10.5	16.1	8.3	12.8
PLP+LDA+MLLT	10.5	16.2	8.4	12.9
PLP+LDA+MLLT+fMLLR	10.5	15.6	8.4	12.2
Posteriorgram combination *V*	**10.0**	**14.9**	**8.1**	**11.7**

†Despite using the same setup and input feature types, there is a mismatch between the results of Chen et al. [12] and our baseline. We believe this is caused by the fact that Chen et al. reportedly use a custom segmentation of the data, where we use the official segmentation of the zero resource challenge 2015. Different segmentations can considerably affect the amount of data actually being used for training.

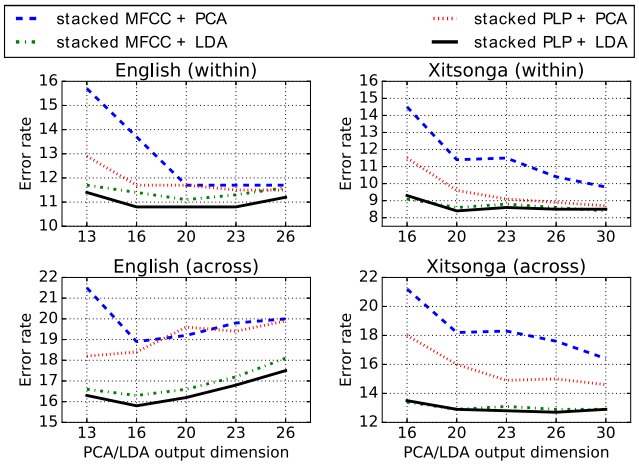

Fig. 3 Discriminability error rates within and across speakers for DPGMM posteriorgrams after clustering PCA or LDA transformed MFCC or PLP feature vectors. The stacking context size is fixed to $c = 4$. The results are plotted as a function of the output dimensionality d of the transformations. *Left:* Error rates for English. *Right:* Error rates for Xitsonga.

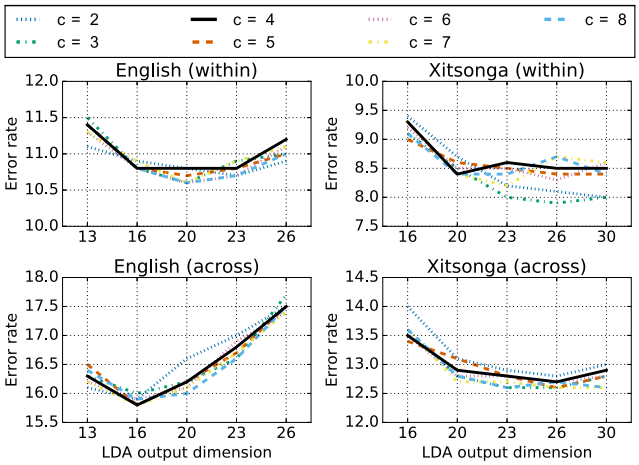

Fig. 4 Discriminability error rates within and across speakers for DPGMM posteriorgrams after clustering LDA transformed PLP feature vectors with varying stacking context size c. The results are plotted as a function of the output dimensionality d of the transformation. *Left:* Error rates for English. *Right:* Error rates for Xitsonga.

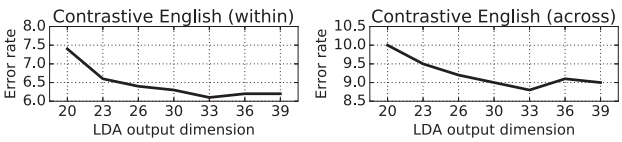

Fig. 5 Discriminability error rates for the contrastive English data set for DPGMM posteriorgrams clustering LDA transformed PLP feature vectors, plotted as a function of the output dimensionality d of the transformations.

Surprisingly, the use of PCA did not show the desired effect of decreasing the discriminability error after DPGMM clustering. In fact, the discriminability error of the GMM posteriorgrams increased on the Xitsonga data. On the English data only little improvement was achieved. The trend is the same whether MFCC or PLP features were transformed.

On the other hand, the LDA transformation produced feature vectors that considerably helped the DPGMM clustering process in finding better clusters, as the discriminability error rates for both data sets decreased greatly, and especially across speakers a strong performance boost is observable. Interestingly, using PLP features for the transformations led to better results than using MFCC features. This is true for both, PCA and LDA transformations.

By using LDA transformed features we already outperform our own baseline and we also beat the numbers of Chen et al. [12]. We take this as proof that the unsupervisedly estimated LDA transformation is the better choice to improve the input to a DPGMM sampler, even when the labels that are used for the estimation are imperfect. The class-discriminating properties of LDA are much more valuable than the simple orthogonalizing that the class-unaware PCA can provide.

6.3 Input and Output Dimensions for LDA

The experiments explained above already show that the choice of the output dimensionality d of transformations influences the clustering performance. We exemplarily conducted a grid search on the parameters c and d LDA transformation of PLP features to find out if this is also true for the input dimensionality. The results of these experiments are visualized in Fig. 4.

The graphs suggest that the impact of the LDA transformation does not depend on the stacking context size c. In our experiments, any context $c > 2$ was suitable. It seems the largest benefit of the dimensional reduction by LDA transformation lies in the compression of the de-correlated features and not so much in the coverage of a larger context.

We see that $d \leq 20$ works well for the English data, and best results for Xitsonga are achieved with $d \geq 20$. We believe this might be due to the data sets' differing speech quality. The English data set consists of conversational speech, and mapping into a lower dimensional space might lead to more stable features for the clustering. The Xitsonga data is read speech, thus the higher dimensions of the transformed feature vectors might still contain distinctive informations.

We conducted additional experiments with a contrastive English data set [27] where we used 38 hours of very clean English read speech to estimate the LDA transformation. Figure 5 shows the error rates on this data set as a function of d. The error curve flattens out in a similar range of values than observed for the Xitsonga data set, which shows comparable speech quality. These results might indicate that d is mainly affected by the quality of the speech.

In a real zero resource scenario we don't have the option to tune c and d. One could therefore try and make an informed guess, or more reasonably use values that have been shown to work well for known languages. We take the latter approach and fix the context to $c = 4$ (i.e., we stack 9 feature vectors) for the input to the LDA, and set the output dimensionality to $d = 20$ (i.e., we keep the first 20 coeffi-

cients), according to the best overall performance on our English data set. By using the parameters we tuned on English, we achieve a performance on Xitsonga that is only slightly lower than the performance that could be achieved with an optimal parameter set, as can be seen in Figs. 3 and 4.

6.4 MLLT

Applying MLLT to LDA transformed features had little to no effect. When we estimate MLLT with our pipeline, the likelihood of the training data is maximized, given the acoustic model that we train along. With the DPGMM, a different generative model for the same data is assumed. Intuitively, it is not guaranteed that MLLT works well in such a cross-model scheme, which our results also show (see Table 1).

6.5 fMLLR

When we applied fMLLR transformations to the feature vector input for the DPGMM sampler, we observed a considerable across speaker discriminability error reduction of the GMM posteriorgrams extracted after the clustering, as seen in Table 1. The relative across speaker error reductions range from 3% to 6%, depending on the data set and the type of the transformed features (MFCC or PLP), but the crucial point is that in our experiments the improvements are independent of data amount, language, and feature types.

Besides doing performance tests, we also analyzed the actual effect of the fMLLR in the feature space. With the frame based class labels from the clustering, we computed the means of the feature vectors for each class and calculated their average distance from each other. We compared the distances of the speaker-dependent means for each class before and after applying fMLLR transformations and found an average distance reduction of 19% and 17% relative for English and Xitsonga. This shows that the fMLLR causes the speaker-dependent means to move closer together, a direct result of removing speaker variance from the features. Interestingly, the speaker-independent means of all the classes moved further away from each other by about 0.7% to 2% relative for English and Xitsonga, and the variance of the features was reduced on average by about 1% relative for both data sets. This means that the fMLLR also helps to increase discriminability between classes. Figure 6 shows the effect of fMLLR in the feature space with an example.

6.6 Posteriorgram Combination

We were using the DPGMM clustering with various kinds of input features and combined the different results with the method from Sect. 4. The expectation was that GMM posteriorgrams from different DPGMM clusterings contain different kinds of latent information about the data and could complement each other in combination.

To produce candidate outputs for combination, we

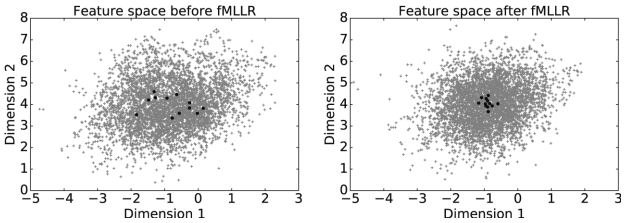

Fig. 6 The figures exemplarily show the 1st and 2nd dimensions of the feature vectors belonging to an arbitrary English acoustic unit as detected by the DPGMM sampler. *Left* is the feature space before, *right* after applying fMLLR transformations. The speaker dependent means (black dots) now cluster in a much smaller area.

Table 2 DPGMMs used for each posteriorgram combination (*Comb.*). The number in brackets behind *LDA* denotes the used output dimension *d*. For combination *V*, eight models were combined, one for each context size *c* between 1 and 8. The context size governs the stacked PLP feature vector size prior to the LDA transformation.

Comb.	#DPGMMs	Clustered features
I	5	PLP+LDA+MLLT+FMLLR
II	3	PLP+LDA PLP+LDA+MLLT PLP+LDA+MLLT+FMLLR
III	2	PLP+LDA+MLLT+FMLLR MFCC+LDA+MLLT+FMLLR
IV	4	PLP+LDA(d = 16)+MLLT+FMLLR PLP+LDA(d = 20)+MLLT+FMLLR PLP+LDA(d = 23)+MLLT+FMLLR PLP+LDA(d = 26)+MLLT+FMLLR
V	8	PLP($1 \leq c \leq 8$)+LDA+MLLT+FMLLR

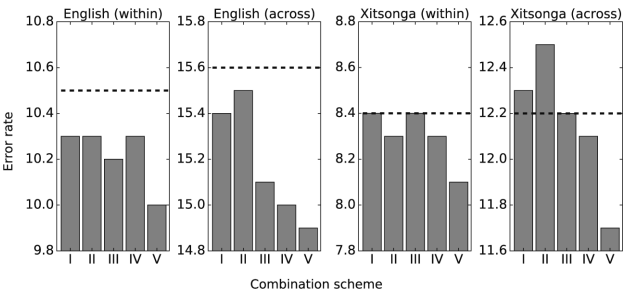

Fig. 7 Discriminability error rates of various posteriorgram combinations. The dotted line marks the best performance on each data set before combining multiple clustering results.

sampled DPGMMs

I multiple times with the same input features,
II for various transformed feature types,
III for transformed MFCC and PLP features,
IV for various LDA output dimensionalities,
V for various LDA input dimensionalities.

Table 2 lists the amount of DPGMMs used in each combination, along with their input features. The discriminability errors of the combined posteriorgrams are plotted in Fig. 7. For the combination experiments we focused on the transformed PLP features, since they generally showed better performance than transformed MFCC features.

Combining the posteriorgrams of 5 DPGMMs that were sampled on the same input features (*I*) only had a

small positive effect on the English data set where the discriminability errors were reduced slightly, compared to the best single DPGMM output. We take this as a sign that the DPGMM sampler generally leads to consistent output, which is why combining results of multiple runs on identical data is particularly helpful. Combination *II* showed similar results.

For combinations *III*, *IV* and *V* we combine the posteriorgrams of DPGMMs that were sampled given more diverse features. The results show that sufficient diversity is critical for the combination to produce better posteriorgrams. In all cases, the combined outputs show lower discriminability errors on the English data set, and can at least match the best single DPGMM output for the Xitsonga data set.

We achieved best results with combination *V*, where we combine posteriorgrams from DPGMMs that were sampled on transformed PLP features with varying context size *c*. The context size governs the stacked PLP feature vector size prior to the LDA transformation. While it seems that an increased context size does not necessarily help the individual DPGMM sampling in particular (as can be seen in Fig. 4), we observed considerable improvements by combining the posteriorgrams produced by these models (see Fig. 7). To ensure that the performance gain is not governed by the choice of the target for the posteriorgram mapping (see Sect. 4), we ran combination *V* multiple times – once for each set of posteriorgrams as target – and averaged the discriminability errors. We found that the average standard deviation across the data is low with 0.05, confirming that the improvements are independent from the choice of the mapping target. The numbers of this best performing combination are found in Table 1, which summarizes our experimental results.

6.7 Analysis

The improvements we have seen after each step in the pipeline are mostly consistent across the two data sets, with the exception of the improvements by LDA (see Table 1). The reductions by fMLLR (0% within and 3.1% to 4.6% across speakers) and by posteriorgram combination (3.5% to 4.7% within and 4% to 4.4% across speakers) are comparable across languages. The improvements by LDA however range from 2.3% to 11% within and 7.8% to 17.8% across speakers, where the larger improvements were observed on English. We believe this is again attributable to the conversational nature of the English data, which provides more room for improvements by LDA. In preliminary experiments on the very clean contrastive English data set mentioned in Sect. 6.3 we observed lower ranges of improvement by LDA (1.5% within and 3% across speakers), which supports our assumption that LDA has more impact on difficult data.

7. Conclusion

We presented a novel approach to optimizing the input of a DPGMM sampler to improve acoustic unit discovery. We evaluated the quality of acoustic unit discovery by computing ABX discriminability errors for posteriorgrams that were extracted from DPGMMs. To substantiate the strengths of our method, we demonstrated its effectiveness on two very different data sets that vary in size, language and speech quality. We demonstrated that it is possible to estimate supervised feature transformations without prior supervision, and that these transformations considerably improve clustering performance. Posteriorgrams of DPGMMs that were sampled given transformed features showed drastically reduced discriminability errors. The use of multiple transformations at once produced better results. A method we introduced for combining the results of multiple DPGMM samplings boosts sound class discriminability even further.

The lowest discriminability errors we achieved are 10% within and 14.9% across speakers for English, and 8.1% within and 11.7% across speakers for Xitsonga. Our proposed framework clearly outperforms our own baseline, as well as the previous state-of-the-art [12]. We believe our approach to optimizing feature vectors for clustering is universal and will be helpful for other zero and low resource tasks as well. In future work we will explore the applicability of our method to other tasks beyond improving automatic unit discovery.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] A. Park and J.R. Glass, "Towards unsupervised pattern discovery in speech," Automatic Speech Recognition and Understanding, IEEE Workshop on, pp.53–58, IEEE, 2005.
[2] A.S. Park and J.R. Glass, "Unsupervised pattern discovery in speech," Audio, Speech, and Language Processing, IEEE Transactions on, vol.16, no.1, pp.186–197, 2008.
[3] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pp.165–168, 2008.
[4] Y. Zhang and J.R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, pp.398–403, IEEE, 2009.
[5] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," Association for Computational Linguistics Annual Meeting, pp.504–511, 2007.
[6] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp.460–470, Association for Computational Linguistics, 2010.
[7] M. Versteegh, R. Thiolliere, T. Schatz, X.N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," Proceedings of Interspeech, pp.3169–3173, 2015.
[8] 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," Proceedings of Interspeech, pp.3199–3203, 2015.
[9] L. Badino, A. Mereta, and L. Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-Markov-model encoders," Proceedings of Interspeech, pp.3174–3178, 2015.
[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," Proceedings of Interspeech, pp.3179–3183, 2015.
[11] H. Kamper, A. Jansen, S. King, and S. Goldwater, "Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings," Spoken Language Technology Workshop (SLT), 2014 IEEE, pp.100–105, IEEE, 2014.
[12] 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," Proceedings of Interspeech, pp.3189–3193, 2015.
[13] R.A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of eugenics, vol.7, no.2, pp.179–188, 1936.
[14] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, pp.661–664, IEEE, 1998.
[15] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," Speech and Audio Processing, IEEE Transactions on, vol.7, no.3, pp.272–281, 1999.
[16] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on, pp.1137–1140, IEEE, 1996.
[17] M.J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer speech & language, vol.12, no.2, pp.75–98, 1998.
[18] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and *k*-means clustering," Proceedings of the 24th international conference on Machine learning, pp.521–528, ACM, 2007.
[19] J. Tang, X. Hu, H. Gao, and H. Liu, "Discriminant analysis for unsupervised feature selection," SDM, pp.938–946, SIAM, 2014.
[20] J. Chang and J.W. Fisher III, "Parallel sampling of DP mixture models using sub-cluster splits," Advances in Neural Information Processing Systems, pp.620–628, 2013.
[21] M.A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Communication, vol.45, no.1, pp.89–95, 2005.
[22] N.J. De Vries, M.H. Davel, J. Badenhorst, W.D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," Speech communication, vol.56, pp.119–131, 2014.
[23] 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," Proceedings of Interspeech, pp.1781–1785, 2013.
[24] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The Kaldi speech recognition toolkit," Proceedings of IEEE, 2011.
[25] K. Pearson, "On lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol.2, no.11, pp.559–572, 1901.
[26] H. Hotelling, "Analysis of a complex of statistical variables into principal components," Journal of educational psychology, vol.24, no.6, pp.417–441, 1933.
[27] E. Dunbar, X.N. Cao, J. Benjumea, J. Karadyi, M. Bernard, L. Besacier, X. Anguerra, and E. Dupoux, "The zero resource speech challenge 2017," Proceedings of ASRU, 2017 (in press).

Michael Heck received his diploma degree in computer science from Karlsruhe Institute of Technology (KIT), Germany, in December 2012. He was granted the Baden-Württemberg scholarship and the Japan Student Services Organization (JASSO) scholarship for a stay at Nara Institute of Science and Technology (NAIST), Japan, during his thesis work. Between 2013-2015, he worked as research assistant at the Interactive Systems Labs at KIT on unsupervised training and adaptation of acoustic models for automatic speech recognition. He was involved in several speech recognition evaluations and international programs including Quaero, Babel and EU-BRIDGE. While at KIT, he was a regular visiting researcher at the Augmented Human Communication Laboratory (AHC Lab) at NAIST. Since 2015, he is a research assistant and doctoral course student at AHC Lab as recipient of the NAIST international scholarship. His research interests include automatic speech recognition, unsupervised learning and zero resource research.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She was actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in 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 speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received 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. 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 Affairs and Communications. He also received LREC Antonio Zampoli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.