SPEAKER INVARIANT FEATURE EXTRACTION FOR ZERO-RESOURCE LANGUAGES WITH ADVERSARIAL LEARNING

Taira Tsuchiya, Naohiro Tawara, Testuji Ogawa and Tetsunori Kobayashi

Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan

ABSTRACT

We introduce a novel type of representation learning to obtain a speaker invariant feature for zero-resource languages. Speaker adaptation is an important technique to build a robust acoustic model. For a zero-resource language, however, conventional model-dependent speaker adaptation methods such as constrained maximum likelihood linear regression are insufficient because the acoustic model of the target language is not accessible. Therefore, we introduce a model-independent feature extraction based on a neural network. Specifically, we introduce a multi-task learning to a bottleneck feature-based approach to make bottleneck feature invariant to a change of speakers. The proposed network simultaneously tackles two tasks: phoneme and speaker classifications. This network trains a feature extractor in an adversarial manner to allow it to map input data into a discriminative representation to predict phonemes, whereas it is difficult to predict speakers. We conduct phone discriminant experiments in Zero Resource Speech Challenge 2017. Experimental results showed that our multi-task network yielded more discriminative features eliminating the variety in speakers.

Index Terms— zero resource speech challenge, speaker invariant feature, adversarial multi-task learning, fMLLR, representation learning

1. INTRODUCTION

There are many languages that have no transcribed data or no written form, and they are called zero-resource languages. Speech processing research for such zero-resource languages has recently attracted increasing attention. For example, a spoken query detection [1, 2], the discovery of sub-word units [3], topic segmentation [4], and document classification [5] have been actively studied for zero-resource languages. In this setting, sub-word units (e.g., phonemes) on the zero-resource languages have to be acquired without any specific knowledge of the target language. Various approaches have been applied to obtain discriminative representations of sub-word units for zero-resource languages. In [6], the sub-word units are automatically generated in nonparametric manner on the target language. Deep neural networks (DNNs) also have been applied to obtain the fine representation of sub-word units, based on sub-word unit classifier, manifold learning [7], and autoencoder [8].

In general, acoustic features have large variations due to the difference in phonemes, noises, channels, and especially speakers. Therefore, the normalization to eliminate the information that does not contribute to distinguish sub-word units has played an important role to build a robust acoustic model. Constrained (feature-space) maximum likelihood linear regression (CMLLR or fMLLR) [9] is widely applied to reduce the variety of speakers. These methods transform the original feature to speaker-invariant one where the likelihood of input sequence against a pre-trained acoustic model is maximized. In the zero-resource setting, however, the acoustic model trained on the target language is generally unaccessible. A simple knowledge transferring approach therefore have been employed in previous researches, where the model trained on non-target rich resource languages (source language) is used for adapting the input data of the target language [6]. These types of model-based adaptation, however, is insufficient when the target language is too far from the source language, especially in the zero-resource scenario.

In this paper, we extended a posteriorgram-based approach with an adversarial learning scheme to enhance the speaker invariance of feature representations. The proposed network is composed of three sub-networks: feature extractor, phoneme classifier, and speaker classifier networks. The representation obtained by the feature extractor is taken as the input to the phoneme and speaker classifier networks. A speaker-invariant representation is obtained by optimizing these models so that phoneme classification error is minimized but speaker classification error is kept high. Our model is tested in the ABX evaluation of the Zero Resource Speech Challenge - Track 1 [10] and yields a certain improvement over the conventional bottleneck approach [11]. Experimental results demonstrate that our model can be transfered to other languages including zero-resource languages. Moreover our approach could be applicable to other neural network based approaches such as autoencoder [8] and Siamese network [7].

2. RELATED WORK

The adversarial learning of DNNs was originally proposed to enhance their robustness against adversarial samples [12].

The adversarial scheme has recently been introduced to classification networks to adapt the input data from one domain to another domain in an unsupervised manner. In this approach, the domain classifier is inserted in an attempt to make the network indiscriminate with respect to the shift between the domains. This framework is applied to document sentiment analysis, image classification, and image re-identification, achieving the high performances [13].

The adversarial learning is also applied in an supervised manner. In [14], adversarial loss was inserted into a senone classification network in an attempt to make the original classification network robust against specific variances. This supervised domain adversarial learning was applied to speech recognition task and showed robustness against different kinds of noise. Following these successes, we applied supervised adversarial learning to eliminate the variety in speakers from bottleneck features for zero-resource languages.

3. SINGLE TASK LEARNING FOR EXTRACTING BOTTLENECK FEATURE

This section briefly explains a conventional supervised neural network for extracting a bottleneck feature with single-task learning. This method is widely used in various applications such as phoneme recognition [15], speaker recognition [15], and language recognition [16, 17].

Let $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ be the training dataset, where K, N, $\boldsymbol{x}_i$, and y_i are the number of classes, number of data, input data, and class label of the i-th data, respectively. The network is trained with the softmax cross entropy loss as follows:

$$P(y_i = k|\boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\exp(a_{ik})}{\sum_{j=1}^K \exp(a_{ij})}, \tag{1}$$

$$L(\boldsymbol{\theta}) = -\frac{1}{n}\sum_i \sum_{k=1}^K y_{ik} \log P(y_i = k|\boldsymbol{x}_i; \boldsymbol{\theta}), \tag{2}$$

where a_{ik} denotes the output of the DNN before taking the softmax function for the k-th output in the i-th data, y_{ik} denotes the one-hot vector of the supervised label y_i, $\boldsymbol{\theta}$ is the set of DNN parameters, and n denotes the size of mini-batch.

Stochastic gradient descent (SGD) [18] is introduced to optimize the above loss function. For each mini-batch, the aforementioned loss function is calculated, and the parameters $\boldsymbol{\theta}$ are updated as follows:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \mu \frac{\partial L}{\partial \boldsymbol{\theta}}, \tag{3}$$

where μ denotes the learning rate. Here, the former part of the DNN plays a role in the feature extraction, and the latter part plays a role as a label predictor. If data is inputted into trained DNN, a certain low-dimensional feature, a bottleneck feature is obtained from the intermediate layer in the DNN.

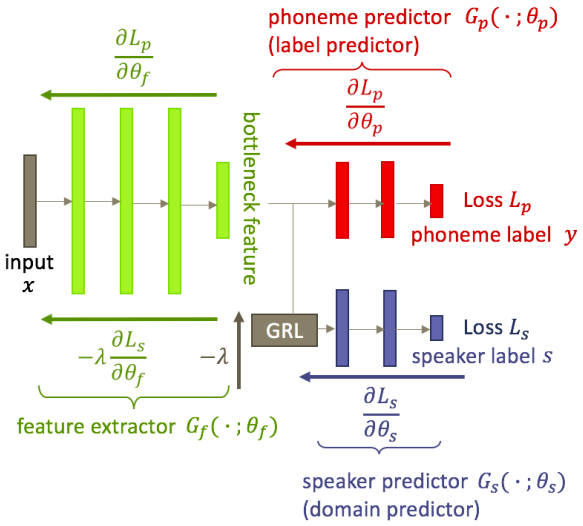


Fig. 1: The structure and flow of loss of proposed adversarial multi-task learning. L_p and L_s correspond to the loss of phoneme and speaker predictors as defined in Eqs. (4) and (5). $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_p$, $\boldsymbol{\theta}_s$ denote the parameters of feature extractor, phoneme predictor, and speaker predictor, respectively. Gradient reversal layer (GRL) acts as an identity transformation in forward propagation and change the sign of loss in back propagation.

4. ADVERSARIAL MULTI-TASK LEARNING FOR SPEAKER INVARIANT BOTTLENECK FEATURE

In this section, adversarial multi-task learning is introduced to supervised neural network described in Section 3 in attempt to eliminate the variety in speakers from the bottleneck feature.

In the adversarial multi-task learning of DNNs, multi kinds of labels are assumed to exist. The sound dataset which has phoneme and speaker labels can be denoted as $\{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^N$, where $s_i \in \{1, ..., C\}$ indicates the speaker class of i-th data, and C is the number of speakers. The difference compared to the single task learning is the use of domain labels for training, with which it is difficult to distinguish speakers for the speaker predictor but easy to discriminate phonemes for the phoneme predictor.

Fig. 1 shows the structure of the adversarial multi-task neural network. This model simultaneously tackles the primary and secondary tasks. In our setting, this model is simultaneously optimized to minimize the loss for phoneme classification (primary task) and speaker prediction (secondary task). The proposed model is composed of three networks: feature extractor, phoneme (label) classifier, and speaker classifier networks. The feature extractor is shared by the primary and secondary tasks and converts the input to bottleneck feature. Then, from the bottleneck feature, the phoneme and the speaker predictors predict phoneme label y and speaker label s, respectively. Here, phoneme prediction loss L_p and speaker

Algorithm 1 Training adversarial multi-task DNN

Input:
- sample $S = \{\boldsymbol{x_i}, y_i, s_i\}_{i=1}^{N}$,
- loss ration parameter λ,
- learning rate μ,

Output: neural network parameter $\{\boldsymbol{\theta}_f, \boldsymbol{\theta}_p, \boldsymbol{\theta}_s\}$

```
while stopping criterion is not met do
  for i from 1 to N do
    # forward propagation
    bottleneck feature f_i ← G_f(x_i; θ_f)
    phoneme output t_i ← G_p(f_i; θ_p)
    speaker output z_i ← G_s(f_i; θ_s)
    # calculate loss
    L_p ← softmax_cross_entropy(t_i, y_i)
    L_s ← softmax_cross_entropy(z_i, s_i)
    # backward propagation
    θ_p ← θ_p − μ ∂L_p/∂θ_p
    θ_s ← θ_s − μ ∂L_s/∂θ_s
    θ_f ← θ_f − μ(∂L_p/∂θ_f − λ ∂L_s/∂θ_f)
  end for
end while
```

prediction loss L_s are denoted as follows:

$$L_p(\boldsymbol{\theta}_f, \boldsymbol{\theta}_p) = -\frac{1}{n}\sum_i \sum_{k=1}^{K} y_{ik} \log P(y_i = k|\boldsymbol{x}_i; \boldsymbol{\theta}_f, \boldsymbol{\theta}_p), \quad (4)$$

$$L_s(\boldsymbol{\theta}_f, \boldsymbol{\theta}_s) = -\frac{1}{n}\sum_i \sum_{c=1}^{C} s_{ic} \log P(s_i = c|\boldsymbol{x}_i; \boldsymbol{\theta}_f, \boldsymbol{\theta}_s), \quad (5)$$

where $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_p$, and $\boldsymbol{\theta}_s$ correspond to the parameters of the feature extractor, the phoneme predictor, and the speaker predictor. Each subscript f, p, s is the first character of each sub-network. n is the size of the mini-batch, and s_{ic} denotes the the c-th value in one-hot expression of label s_i.

Algorithm 1 denotes the pseudo code of the training procedure of the proposed adversarial multi-task DNN. The aim of this model is to extract features that are easy for phoneme predictor to classify phonemes and difficult for speaker predictor to classify speakers. As a result, speaker-independent bottleneck features are obtained. The most critical parameter update used for feature extractor parameters is as follows:

$$\boldsymbol{\theta}_f \leftarrow \boldsymbol{\theta}_f - \mu(\frac{\partial L_p}{\partial \boldsymbol{\theta}_f} - \lambda \frac{\partial L_s}{\partial \boldsymbol{\theta}_f}), \quad (6)$$

where λ means the extent that the feature extractor cannot extract features that are easy to determine speakers, and μ is the learning rate. Note that feature extractor of the single task learning is identical to that of the multi-task learning with $\lambda = 0$.

5. EXPERIMENTS

5.1. Dataset

Experimental comparisons were carried out using English, French and Mandarin contained in the Zero Resource Speech Challenge 2017 [10]. In these experiments, English was used as the source language, and French and Mandarin were used as the zero-resource languages. Note that the acoustic models for French and Mandarin could not be constructed and accessible since these languages are regraded as zero-resource languages. An acoustic model trained on English dataset was adopted to obtain fMLLR features of the zero-resource languages. Both single-task and multi-task networks were trained with English data spoken by nine speakers. The duration of each speaker's utterance was ranging from 165 min to 220 min. The total duration of English, French, and Mandarin for test sets were 1634 min, 1061 min, and 1522 min, respectively. Speech data for test set were segmented into 120 seconds. Although English was not regarded as a zero-resource language, but a test for English was also conducted to validate the effectiveness of the proposed method in resource-abundant languages.

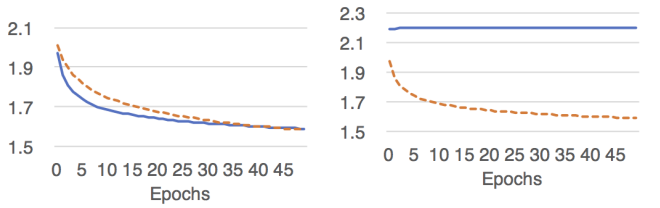
We firstly removed silence region. 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their Δ parameters were extracted with a 25 ms analysis window and 10 ms window shift, followed by mean and variance normalization (MVN) to each segment. Then, each frame of target language is linearly transformed with transformation matrix trained on source language (English) and 40-dimensional features are obtained. Note that this linear transformation is obtained by discriminant analysis, maximum likelihood linear transformation, and feature space MLLR trained on source language (English). The 40 dimension fMLLR feature is concatenated with five frames before and after. Thus, a 440 dimension fMLLR feature was used as the input.

5.2. Experimental settings

The models were implemented with Chainer [19]. The numbers of units in each layer of the feature extractor, phoneme, and speaker predictors were {440, 1024, 1024, 1024, 256}, {256, 256, 40}, and {256, 256, 9}, respectively. The mini-batch size was set to 1024. Dropout [20] and batch normalization [21] layers were inserted after each hidden layer to prevent over-fitting. The dropout ratio for the feature extractor, phoneme, and speaker predictors were set to 0.1, 0.2, and 0.2, respectively. All models were trained with SGD with learning rate decay as follows:

$$\mu_p = \frac{\mu_0}{(1 + \alpha \cdot p)^{\beta}}, \quad (7)$$

where p denotes the training progress, which increases linearly from zero to one in the same manner as was done in [13]. We set $\mu_0 = 0.01$, $\alpha = 10$, $\beta = 0.75$.



(a) single-task learning

(b) adversarial multi-task learning ($\lambda = 5$)

Fig. 2: Validation losses obtained in (a) single-task learning and (b) adversarial multi-task learning. Dashed and solid lines depict phoneme and speaker prediction losses, respectively.

5.3. Evaluation metric

Performances of the models were evaluated by measuring the phoneme discriminability of obtained bottleneck features for low-resource languages. ABX discriminability [22] was adopted for the evaluation. This metrics is based on an ABX task, where we test which of a minimal pair of sound (A and B) belonged to the same category as given sound X. If X belongs to A, the distance $d(A, X)$ between A and X should be smaller than that between B and X over the sound feature space. Based on this assumption, the ABX error rate is calculated using the sets of each phonemes, $S(\boldsymbol{x})$, $S(\boldsymbol{y})$ as:

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{m(m-1)n} \sum_{a \in S(\boldsymbol{x})} \sum_{b \in S(\boldsymbol{y})} \sum_{x \in S(\boldsymbol{x}) \setminus \{a\}} \left(\mathbb{1}_{d(a,x) < d(b,x)} + \frac{1}{2} \mathbb{1}_{d(a,x) = d(b,x)} \right) \tag{8}$$

where m and n correspond to the number of examples that belong to $S(\boldsymbol{x})$ and $S(\boldsymbol{y})$, and $\mathbb{1}$ is an indicator function. The distance between sounds $\boldsymbol{x}$ and $\boldsymbol{y}$, $d(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ can have different length, was calculated by the dynamic time warping on the underlying frame-to-frame distance with cosine similarity.

5.4. Results

Fig. 2 demonstrates the validation losses obtained in each epoch of training: (a) single and (b) adversarial multi-task learning. This figure demonstrates that, in the single task learning, both the phoneme and the speaker classification losses decreased as the number of epochs increased. On the other hand, in the adversarial multi-task learning, only the phoneme classification loss decreased, while speaker classification loss increased. This result indicated that the adversarial learning on the speaker classification network correctly worked and provided the speaker invariant feature.

Tables 1 and 2 list the ABX error values across and within the speakers conditions, respectively. Here, we can see that all the results that used the fMLLR feature outperformed those that used MFCCs by a large margin. This result indicates that the fMLLR decoded by English can be applicable to the other languages. Second, the bottleneck feature outperformed for both the MFCCs and fMLLR features. Finally, the proposed adversarial multi-task learning (AMTL) outperformed the single task learning (STL) for all the languages and achieved the best result of all the methods. This indicates that AMTL could reduce the variation of the speakers.

Table 1: ABX error rate across speakers' test data for zero resource languages (French, Mandarin) and resource-abundant language (English). The Baselines directly used the features (MFCCs and fMLLR). STL and AMTL represent the single-task learning and adversarial multi-task learning, respectively.

	English	French	Mandarin
Baseline (MFCCs)	23.4	25.2	21.3
Baseline (fMLLR)	10.832	14.832	10.351
STL (fMLLR)	7.064	12.100	8.901
AMTL (fMLLR)	**6.796**	**11.866**	**8.725**

Table 2: ABX error rate for within speakers test data

	English	French	Mandarin
Baseline (MFCCs)	12.1	12.6	11.5
Baseline (fMLLR)	6.846	8.955	8.740
STL (fMLLR)	4.957	8.060	8.008
AMTL (fMLLR)	**4.707**	**7.592**	**7.819**

6. CONCLUSION

This paper introduced a novel approach to extract a speaker invariant feature for zero-resource languages. Adversarial multi-task learning was introduced to make a bottleneck feature invariant to a change of speakers. ABX experiments was conducted on one known language and two unknown languages. The experimental comparison demonstrated that the proposed adversarial multi-task learning outperformed the conventional single-task approach under all conditions.

The advantage of the proposed multi-task learning is that its applicability to various types of DNNs. We therefore plan to apply the proposed multi-task learning to various kinds of neural networks such a bottleneck approach based on multilingual phonemes [23] and a Siamese network [7]. We also plan to apply our model to the Dirichlet process GMM approach for zero resource languages [6, 24]. Meanwhile, we have still question whether we could actually extract the characteristics of a speaker from acoustic feature within acoustic feature $20\text{ms} \times 11 = 220\text{ms}$. Thus, we plan to extend our model to use information in a longer context by using a long short time memory network [25] or i-vector [26].

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP16K12465.

7. REFERENCES

[1] Gautam Mantena and Kishore Prahallad, "Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios," in *ICASSP*, 2014, pp. 7128–7132.

[2] Yaodong Zhang and James R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *ASRU*, 2009, pp. 398–403.

[3] Chia-ying Lee and James Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *ACL*, 2012, vol. 1, pp. 40–49.

[4] Park A. Barzilay R. Glass J. Malioutov, I., "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *ACL*, 2007, vol. 45, pp. 504–511.

[5] Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church, "NLP on spoken documents without ASR," in *EMLNLP*, 2010, pp. 460–470.

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

[7] Roland Thiollire, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *INTERSPEECH*, 2015, pp. 3179–3183.

[8] Leonardo Badino, Alessio Mereta, and Lorenzo Rosasco, "Discovering discrete subword units with Binarized Autoencoders and Hidden-Markov-Model Encoders," in *INTERSPEECH*, 2015, pp. 3174–3178.

[9] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[10] "The Zero Resource Speech Challenge 2017," "http://sapience.dec.ens.fr/bootphon/2017/", Accessed on 2017-10-27.

[11] Daniel Renshaw andHerman Kamper andAren Jansen andSharon Goldwater, "A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge," in *INTERSPEECH*, 2015, pp. 3199–3203.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 1, pp. 2096–2030, 2016.

[14] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *INTERSPEECH*, 2016, pp. 2369–2372.

[15] Sibel Yaman, Jason Pelecanos, and Ruhi Sarikaya, "Bottleneck features for speaker recognition," in *IEEE Odyssey*, 2012.

[16] Pavel Matejka, Le Zhang, Tim Ng, Harish Sri Mallidi, Ondrej Glembek, Jeff Ma, and Bing Zhang, "Neural network bottleneck features for language identification," *Proc. IEEE Odyssey*, pp. 299–304, 2014.

[17] Fred Richardson, Douglas Reynolds, and Najim Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[18] Léon Bottou, "Stochastic gradient learning in neural networks," *Proc. of Neuro-Nımes*, vol. 91, no. 8, 1991.

[19] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton, "Chainer: a next-generation open source framework for deep learning," in *Workshop on Machine Learning Systems in NIPS*, 2015.

[20] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[21] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, vol. 37, pp. 448–456.

[22] Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *INTERSPEECH*, 2013, pp. 1–5.

[23] R. Fér, P. Matějka, F. Grézl, O. Plchot, K. Veselỳ, and J. H. Černockỳ, "Multilingually trained bottleneck features in spoken language recognition," *Computer Speech & Language*, vol. 46, pp. 252–267, 2017.

[24] Michael Heck, Sakriani Sakti, and Satoshi Nakamura, "Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering," in *INTERSPEECH*, 2016, pp. 1310–1314.

[25] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] Najim Dehak, Patrick J. Kenny, Rda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.