ROLE PLAY DIALOGUE TOPIC MODEL FOR LANGUAGE MODEL ADAPTATION IN MULTI-PARTY CONVERSATION SPEECH RECOGNITION

Ryo Masumura, Takanobu Oba, Hirokazu Masataki, Osamu Yoshioka, Satoshi Takahashi

NTT Media Intelligence Laboratories, NTT Corporation, Japan
{masumura.ryo, oba.takanobu, masataki.hirokazu, yoshioka.osamu, takahashi.satoshi}@lab.ntt.co.jp

ABSTRACT

This paper introduces an unsupervised language model adaptation technique for multi-party conversation speech recognition. The use of topic models provides one of the most accurate frameworks for unsupervised language model adaptation since they can inject long-range topic information into language models. However, conventional topic models are not suitable for multi-party conversation because they assume that each speech set has each different topic. In a multi-party conversation, each speaker will share the same conversation topic and each speaker utterance will depend on both topic and speaker role. Accordingly, this paper proposes new concept of the "role play dialogue topic model" to utilize multi-party conversation attributes. The proposed topic model can share the topic distribution among each speaker and can also consider both topic and speaker role. The proposed topic model based adaptation realizes a new framework that sets multiple recognition hypotheses for each speaker and simultaneously adapts a language model for each speaker role. We use a call center dialogue data set in speech recognition experiments to show the effectiveness of the proposed method.

Index Terms— Unsupervised language model adaptation, multi-party conversation speech recognition, topic model

1. INTRODUCTION

With the recent development of automatic speech recognition technology, multi-party conversation tasks such as contact center dialogue or meeting have been attracting attention [1, 2, 3]. Multi-party conversation tasks have different attributes from the typical single speaker task such as lecture speech, since several speakers interact with each other. As this interaction can infect each speaker utterance, it is necessary to develop a new speech recognition framework that can take account of this interaction. Thus this paper focuses on unsupervised language model (LMs) adaptation for multi-party conversation speech recognition [4].

Several techniques have been proposed for unsupervised LM adaptation [5]. One of the most accurate approaches for unsupervised LM adaptation is based on probabilistic topic models. Topic models such as Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA) can capture the semantic properties of words and documents [6, 7]. While n-gram LMs suffer from insufficient long-range information, topic models can capture long-range topic information. In the case of unsupervised LM adaptation using topic models, the topic probability of the target speech is estimated using recognition hypothesis with the result being adapted unigram probability. Next, the n-gram LM is adapted based on unigram marginal [8, 9] or linear interpolation [10, 11].

However, conventional topic models are not suitable for multi-party conversation because they assume that each speech set has a different topic. Briefly, conventional topic models are only appropriate for the single speaker task. In multi-party conversation, we have to give consideration to the aspect of the correlation among speech sets. We assume that multi-party conversations have the following attributes. First, each speaker shares the same conversation topic. Moreover, each speaker utterance will depend on not only conversational topic but also own role. The role means speaker type. For example, there are two roles in contact center dialogue, which are operator and customer. Although topic models with a topic sharing structure were proposed in the machine translation area [12, 13, 14, 15], they are not suitable for multi-party conversations because they cannot take account of speaker role and only deal with language translation problems.

Our solution is to propose a new topic model that can utilize multi-party conversation attributes. The proposed topic model shares the topic distribution among all speakers and also considers both topic and speaker role. The proposed topic model based unsupervised LM adaptation realizes a new framework that sets multiple recognition hypotheses for each speaker and simultaneously adapts LMs for each speaker role using the shared conversation topic. As the adaptation proposal can take both topic sharing and speaker role into consideration, we can expect further improvements in ASR performance. In fact, considering the participation of several speakers in the same conversation is effective for language modeling [16]. Furthermore, considering speaker type is also beneficial for language model adaptation [17].

The rest of this paper is organized as follows. First, the

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

new topic model is briefly described in Section 2. Section 3 explains unsupervised LM adaptation using the proposed topic models. Section 4 describes our experiments and results. Finally, our conclusion is described in Section 5.

2. ROLE PLAY DIALOGUE TOPIC MODEL

2.1. Definition

We propose new topic models called role play dialogue topic models (RPDTMs) for multi-party conversations. RPDTM assumes that each speaker role is given and each speaker's utterance is divided. This situation can be duplicated in contact center dialogue or poster presentation.

The model definition is as follows. Topic index is represented as $k \in \{1, \cdots, K\}$. Speaker role index is represented as $r \in \{1, \cdots, R\}$, and dialogue index is represented as $m \in \{1, \cdots, M\}$. Dialogue data about m, which includes several speaker's data, is shown as $\boldsymbol{D}_m = \{\boldsymbol{W}_{1,m}, \cdots, \boldsymbol{W}_{R,m}\}$. RPDTM also assumes that dialogue data set $\boldsymbol{O} = \{\boldsymbol{D}_1, \cdots, \boldsymbol{D}_M\}$ is generated according to the following stochastic process.

1. For each topic $k = 1, ..., K$:
   (a) Draw $\delta_k \sim Dirichlet(\beta)$
2. For each speaker role $r = 1, ..., R$:
   (a) Draw $\phi_r \sim Dirichlet(\delta)$
   (b) Draw $\pi_r \sim Dirichlet(\gamma)$
3. For each dialogue $m = 1, ..., M$:
   (a) Draw $\theta_m \sim Dirichlet(\alpha)$
   (b) For each speaker role $r = 1, ..., R$
      i. For each word $i = 1, ..., N_{r,m}$:
         A. Draw $z_i \sim Multinomial(\theta_m)$
         B. Draw $c_i \sim Multinomial(\pi_r)$
         C. If $c_i = 1$,
            then draw $w_i \sim Multinomial(\varphi_{z_i})$,
            else draw $w_i \sim Multinomial(\phi_r)$

where $N_{r,m}$ indicates the number of words in $\boldsymbol{W}_{r,m}$. *Dirichlet* means a Dirichlet distribution, and *Multinomial* means a multinomial distribution. α, β, γ, δ are hyper parameters for each Dirichlet distribution. θ, π, φ, ϕ are model parameters for each multinomial distribution. z is a topic variable, and c is a condition variable. c controls whether a certain word is dependent on dialogue topic or speaker role. A graphic rendering of RPDTM is shown in Figure 1.

Next, we introduce equations to detail the generation of dialogue data sets. The generation probability of dialogue data set $\boldsymbol{O}$ is given by:

$$P(\boldsymbol{O}|\boldsymbol{\Theta}) = \prod_{m=1}^{M} \int P(\theta_m|\alpha) \prod_{r=1}^{R} P(\boldsymbol{W}_{r,m}|\boldsymbol{\Theta}) d\theta_m, \quad (1)$$

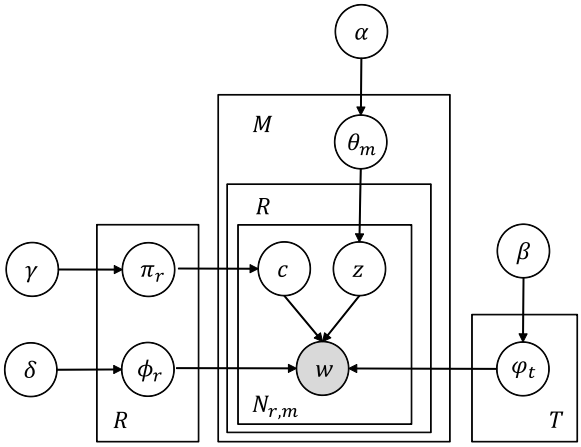

Fig. 1. Model strucure of RPDTM.

where $\boldsymbol{\Theta}$ are hyper parameters and other parameters generated in the middle of the process. Eq. (1) is similar to the generation probability of data sets in LDA [7]. In LDA, a topic distribution is generated for each data. On the other hand, RPDTM generates a topic distribution for each dialogue, which includes several speech sets.

The word sequence associated with speaker role r in dialogue m is represented as $\boldsymbol{W}_{r,m} = \{w_1, \cdots, w_{N_{r,m}}\}$. the generation probability of $\boldsymbol{W}_{r,m}$ is given by:

$$P(\boldsymbol{W}_{r,m}|\boldsymbol{\Theta}) = \prod_{i=1}^{N_{r,m}} P(w_i|\boldsymbol{\Theta}). \quad (2)$$

RPDTM is a bag of words model, so each word is generated independently in common with LDA. The generation probability for each word w is shown as:

$$P(w|\boldsymbol{\Theta}) = \sum_{c} P(w|c, \boldsymbol{\Theta}) P(c|\pi_r, \gamma). \quad (3)$$

In RPDTM, $P(w|c, \boldsymbol{\Theta})$ has different form depending on condition variable c. c takes the values of 0 or 1. If $c = 0$, it is probable that $P(w|c, \boldsymbol{\Theta})$ is related to speaker role. On the other hand, if $c = 1$, it is likely to be related to the topic of the dialogue. $P(w|c, \boldsymbol{\Theta})$ is given by Eq. (4).

$$P(w|c, \boldsymbol{\Theta}) = \begin{cases} P(w|\phi_r, \delta) & (c = 0), \\ \sum_{z} P(w|\varphi_z, \beta) P(z|\theta_m, \alpha) & (c = 1). \end{cases} \quad (4)$$

RDPTM can simultaneously treat both topic and speaker role. HMM-LDA uses a similar generation process to treat both topic and context information [18].

2.2. Inference

Inference in RPDTM means estimating topic variable assignments and condition variable assignments of all words in the

training data. Once topic variable assignment and condition variable assignment are concluded, generation probabilities of each variable are calculated as follows:

$$P(z|\theta_m, \alpha) = \frac{\sum_r n_{r,m}(z) + \alpha}{\sum_r N_{r,m} + K\alpha}, \tag{5}$$

$$P(c|\pi_r, \gamma) = \frac{\sum_m n_{r,m}(c) + \gamma}{\sum_m N_{r,m} + 2\gamma}, \tag{6}$$

where $n_{r,m}(z)$ is the number of words assigned to topic variable z, and $n_{r,m}(c)$ is the number of words assigned to condition variable c associated with speaker role r in dialogue m.

Next, generation probabilities of word w are defined as follows:

$$P(w|\varphi_z, \beta) = \frac{n_0(w, z) + \beta}{n_0(z) + V\beta}, \tag{7}$$

$$P(w|\phi_r, \delta) = \frac{\sum_m n_{r,m}(w, c) + \delta}{\sum_m n_{r,m}(c) + V\delta}, \tag{8}$$

where V represents vocabulary size. $n_{r,m}(w, c)$ means the number of times word w is assigned to condition variable c associated with speaker role r in dialogue m. $n_0(z)$ means the number of words assigned to topic variable z and condition variable 1 in dialogue data set, and $n_0(w, z)$ means the number of times word w is assigned to topic variable z and condition variable 1 in the dialogue data set.

Gibbs sampling is used for the assignment of topic variable and condition variable [19]. Here, a word sequence associated with speaker role r in dialogue m is expressed as $\boldsymbol{W} = \{w_1, \cdots, w_{N_{r,m}}\}$, and the corresponding topic variable sequence and condition variable sequence are expressed as $\boldsymbol{Z} = \{z_1, \cdots, z_{N_{r,m}}\}$ and $\boldsymbol{C} = \{c_1, \cdots, c_{N_{r,m}}\}$, respectively. A conditional probability of possible values for topic variable z_i is obtained as:

$$P(z_i|\boldsymbol{Z}^{-i}, \boldsymbol{W}, \boldsymbol{C}, \boldsymbol{U}^{-(r,m)}) \propto \begin{cases} P(z_i|\theta_m, \alpha) & (c_i = 0), \\ P(z_i|\theta_m, \alpha)P(w_i|\varphi_{z_i}, \beta) & (c_i = 1), \end{cases} \tag{9}$$

where $\boldsymbol{Z}^{-i}$ represents all latent variables except for z_i, and $\boldsymbol{U}^{-(r,m)}$ denotes all of the parameter set except for data about speaker role r in dialogue m.

Next, a conditional probability of possible values for condition variable c_i is obtained as:

$$P(c_i|\boldsymbol{C}^{-i}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{U}^{-(r,m)}) \propto \begin{cases} P(c_i|\pi_r, \gamma)P(w_i|\phi_r, \delta) & (c_i = 0), \\ P(c_i|\pi_r, \gamma)P(w_i|\varphi_{z_i}, \beta) & (c_i = 1), \end{cases} \tag{10}$$

where $\boldsymbol{C}^{-i}$ represents all condition variables except for c_i. Gibbs sampling can be used to sample a new value for the topic variable and condition variable according to these two distributions and place it at position i.

3. PROPOSED UNSUPERVISED LM ADAPTATION

Unsupervised LM adaptation for automatic speech recognition uses the concept of recognition hypotheses. In common with conventional topic model based adaptation, topic probability is estimated using recognition hypotheses and adapted unigram probability is calculated. Next, n-gram LM is adapted using adapted unigram probability [8, 9, 10, 11].

We assume that speech recognition of a multi-party conversation is performed for every speaker role. The proposed unsupervised LM adaptation simultaneously uses all recognition hypotheses to recognize the target dialogue. We use RPDTM to estimate topic probability for the target dialogue and simultaneously calculate unigram probabilities for each speaker role.

Here, we define $\bar{m}$ as the index of the recognition target dialogue. To estimate the topic probability of the target dialogue, Gibbs sampling is used. In this case, we have to estimate topic variable assignments and condition variable assignments of all words in the recognition hypotheses. These procedures are based on Eq. (9) and Eq. (10). If each variable assignment is defined, we can calculate topic probability $P(z|\theta_{\bar{m}}, \alpha)$ based on Eq. (6). Then, adapted unigram probabilities $P_{1,\bar{m}}(w), \cdots, P_{R,\bar{m}}(w)$ are obtained by assigning $P(z|\theta_{\bar{m}}, \alpha)$ to Eq. (4).

The adaptation of n-gram LMs also considers each speaker role. This paper uses the unigram marginal technique, which can consider back-off probabilities in n-gram probability [20, 21]. Adapted n-gram LM probability of word w in given context $\boldsymbol{u}$ is defined as follows:

$$P_{r,\bar{m}}(w|\boldsymbol{u}) = \frac{\tau_{r,\bar{m}}(w)P_0(w|\boldsymbol{u})}{Z(\boldsymbol{u})}, \tag{11}$$

where $P_{r,\bar{m}}(w|\boldsymbol{u})$ is n-gram LM adapted with regard to speaker role r in target dialogue $\bar{m}$, and $P_0(w|\boldsymbol{u})$ is the background n-gram LM constructed from training data. $Z(\boldsymbol{u})$ is a normalization term and is given by:

$$Z(\boldsymbol{u}) = \sum_w \tau_{r,\bar{m}}(w)P_0(w|\boldsymbol{u}). \tag{12}$$

$\tau_{r,\bar{m}}(w)$ is a scaling factor that is defined as follows:

$$\tau_{r,\bar{m}}(w) \approx \left(\frac{P_{r,\bar{m}}(w)}{P_0(w)}\right)^{\mu}, \tag{13}$$

where μ is tuning parameter, and $P_0(w)$ is ML unigram probability estimated from training data. $P_{r,\bar{m}}(w)$ is unigram probability about speaker role r in target dialogue $\bar{m}$, which is determined by RPDTM from the recognition hypotheses.

4. EXPERIMENTS

4.1. Experimental conditions

Our experiments used the contact center dialogue data sets, which include several topics. One dialogue set means one

Table 2. *Experimental results.*

	# of topics	Topic sharing	Consideration for speaker role	Dev. PPL	Dev. WER (%)	Test A PPL	Test A WER (%)	Test B PPL	Test B WER (%)
1. BASE	-	-	-	33.77	21.37	46.55	24.69	47.12	22.70
2. LDA1	20	×	◯	31.43	21.15	41.73	24.35	42.56	22.26
2. LDA1	30	×	◯	31.55	20.94	41.96	24.55	42.44	22.44
3. LDA2	20	◯	×	32.04	21.06	42.98	25.20	43.88	22.18
3. LDA2	30	◯	×	31.90	21.12	42.18	24.63	44.15	22.36
4. RPDTM	20	◯	◯	29.82	20.61	**37.91**	**23.46**	**39.66**	**21.20**
4. RPDTM	30	◯	◯	**29.57**	**20.45**	38.25	23.65	39.96	21.36

Table 1. *Experimental data set.*

	# of dialogues	# of words
Training	1,922	1,659,230
Development	8	8,277
Test A	8	7,393
Test B	8	8,995

telephone call between one operator and one customer. Each dialogue was separately recorded and the data set consists of 1,984 dialogues. We divided this data set into a training set, a development set, and a test set. Table 1 shows details.

We used an acoustic model based on hidden Markov models with deep neural networks (DNN-HMM) [22]. DNN-HMM was trained with 7 hidden layers of 2048 nodes and 3874 outputs. The speech recognition decoder is VoiceRex, a WFST-based decoder [23, 24]. JTAG was used as the morpheme analyzer to split sentences into words [25]. We constructed a 3-gram hierarchical Pitman-Yor LM as the background n-gram LM [26]. Vocabulary size was 60K.

For evaluation, we compared RPDTM-based adaptation to conventional LDA-based adaptation [9]. The comparison used the following methods.

1. **BASE**: First pass based on background n-gram LM.
2. **LDA1**: Individually construct adapted LM using each speaker recognition hypothesis based on LDA. Although this method can realize speaker role dependent adaptation, topic sharing is not considered.
3. **LDA2**: Construct single adapted LM using all speaker recognition hypotheses based on LDA. Although this method can perform topic sharing, it cannot take speaker role into consideration.
4. **RPDTM**: Individually construct adapted LM using all speaker recognition hypotheses based on RPDTM. This method can take both topic sharing and speaker role into consideration.

LDA and RPDTM were formed on the training data. We used 500 iterations for Gibbs sampling. α, β, γ and δ, the hyper parameters of RPDTM, were set to 0.5. Tuning parameter for unigram marginal was set to 0.6 as this value was found to be to optimal for the development set. The number of topics in LDA and RPDTM was set to 20 or 30.

4.2. Experimental results

We investigated the perplexity (PPL) and word error rate (WER) results for the development set and each test set. The results, shown in Table 2, confirm that topic model based unsupervised LM adaptation is especially effective in terms of PPL, but WER improvements are difficult to see. Regardless of the number of topics, the best results for the development set and each test set were obtained by RPDTM-based adaptation. LDA1 and RPDTM differ in terms of whether to share the topic, and it seems that topic sharing is effective. Moreover, LDA2 and RPDTM differ in terms of whether LMs are individually adapted for each speaker role, and it seems that role dependent adaptation is effective. RPDTM-based adaptation can take both topic sharing and role-dependent adaptation into consideration, simultaneously. It seems that unsupervised language model adaptation is most effective if it utilizes dialog features.

5. CONCLUSIONS

In this paper, we proposed a topic model and unsupervised LM adaptation for multi-party conversation speech recognition. The proposed model, RPDTM, shares the topic distribution among each speaker and can also consider both topic and speaker role. The proposed topic model based adaptation realizes a new framework that sets multiple recognition hypotheses for each speaker and simultaneously adapts a language model for each speaker role.

Experiments showed that RPDTM-based adaptation is more effective than LDA-based adaptation, confirming that both topic sharing and role-dependent adaptation must be taken into consideration when performing LM adaptation for multi-party conversations.

We plan to conduct experiments on the meeting task in future research. Furthermore, we intend to realize a dynamic adaptation framework, such as a topic tracking language model [27], for multi-party conversation.

6. REFERENCES

[1] L. Venkata Subramaniam, Tanveer A. Faruquie, Shajith Ikbal, Shantanu Godbole, and Mukesh K. Mohania, "Business intelligence from voice of customer," *In Proc. ICDE*, pp. 1391–1402, 2009.

[2] Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke, "The meeting project at icsi," *In Proc. HLT*, pp. 1–7, 2001.

[3] Steve Renals, Thomas Hain, and Herve Bourlard, "Recognition and understainding of meetings the ami and amida project," *In Proc. ASRU*, pp. 238–247, 2007.

[4] Gokhan Tur and Andreas Stolcke, "Unsupervised language model adaptation for meeting recognition," *In Proc. ICASSP*, vol. 4, pp. 173–176, 2007.

[5] Michiel Bacchiani and Brian Roark, "Unsupervised language model adaptation," *In Proc. ICASSP*, pp. 224–227, 2003.

[6] Thomas Hofmann, "Probabilistic latent semantic analysis," *In Proc. UAI*, pp. 289–296, 1999.

[7] David M Blei, Andrew Y Ng, and Micheal I Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Reseach*, pp. 993–1022, 2003.

[8] Marcello Federico, "Language model adaptation through topic decomposition and mdi estimation," *In Proc. ICASSP*, vol. 1, pp. 703–706, 2002.

[9] Yik-Cheung Tam and Tanja Schultz, "Unsupervised language model adaptation using latent semantic marginals," *In Proc. Interspeech 2006*, pp. 2207–2209, 2006.

[10] David Mrva and Philip C. Woodland, "A plsa-based language model for conversational telephone speech," *In Proc. ICSLP*, 2004.

[11] Yik-Cheung Tam and Tanja Schultz, "Dynamic language model adaptation using variation bayes inference," *In Proc. ICASSP*, pp. 5–8, 2005.

[12] Yik-Cheung Tam and Tanja Schultz, "Bilingual-lsa based lm adaptation for spoken language translation," *In Proc. ACL*, pp. 520–527, 2007.

[13] Zhengxian Gong, Yu Zhang, and Guodong Zhou, "Statistical machine translation based on lda," *In Proc. IUCS*, pp. 286–290, 2010.

[14] David Mimno, Hanna M. Wallach, Jacson Naradowsky, David A. Smith, and Andrew Mcllum, "Polylingual topic model," *In Proc. EMNLP*, pp. 880–889, 2009.

[15] Nick Ruiz and Marcello Federico, "Topic adaptation for lecture translation through bilingual latent semantic models," *In Proc. Sixth Workshop in Statistical Machine Translation*, pp. 294–302, 2011.

[16] Gang Ji and Jeff Bilmes, "Multi-speaker language modeling," *In Proc. HLYNACAC*, pp. 133–136, 2004.

[17] Yik-Cheung Tam and Paul Vozila, "Unsupervised latent speaker language modeling," *In Proc. Interspeech 2011*, pp. 1477–1480, 2011.

[18] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum, "Integrating topics and syntax," *In Proc. Neural Information Processing Systems*, pp. 537–544, 2004.

[19] George Casella and Edward I George, "Explaining the gibbs sampler," *The American Statistician*, vol. 46, pp. 167–174, 1992.

[20] Reinhard Kneser, Jochen Peters, and Dietrich Klakow, "Language model adaptation using dynamic marginals," *In Proc. Eurpspeech*, pp. 1971–1974, 1997.

[21] Marcello Federico, "Efficient language model adaptation through mdi estimation," *In Proc. Eurospeech*, vol. 4, pp. 1583–1586, 1999.

[22] Frank Seide, Gang Li, and Dong Yu, "Coversational speech transcription using context-dependent deep neural networks," *In Proc. Interspeech 2011*, pp. 437–440, 2011.

[23] Takaaki Hori, Chiori Hori, Yasuhiro Minami, and Atsushi Nakamura, "Efficient wfst-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.

[24] Hirokazu Masataki, Daisuke Shibata, Yuichi Nakazawa, Satoshi Kobashikawa, Atsunori Ogawa, and Katsutoshi Ohtsuki, "Voicerex spontaneous speech recognition technology for contact-center conversations," *NTT Tech. Rev.*, vol. 5, no. 1, pp. 22–27, 2007.

[25] Takeshi Fuchi and Shinichiro Takagi, "Japanese morphological analyzer using word co-occurence-jtag," *In Proc. COLING-ACL*, pp. 409–413, 1998.

[26] Yee Whye Teh, "A hierarchical bayesian language model based on pitman-yor processes," *In Proc. COLING/ACL 2006*, pp. 985–992, 2006.

[27] Shinji Watanabe, Tomoharu Iwata, Takaaki Hori, Atsushi Sako, and Yasuo Ariki, "Topic tracking language model for speech recognition," *Computer Speech & Language*, vol. 25, pp. 440–461, 2011.