INTERSPEECH 2009 BRIGHTON

Rapid Unsupervised Adaptation Using Frame Independent Output Probabilities of Gender and Context Independent Phoneme Models

KOBASHIKAWA Satoshi[1], *OGAWA Atsunori*[2], *YAMAGUCHI Yoshikazu*[1], *TAKAHASHI Satoshi*[1]

[1]NTT Cyber Space Laboratories, NTT Corporation, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan

10.21437/Interspeech.2009-211

Abstract

Business is demanding higher recognition accuracy with no increase in computation time compared to previously adopted baseline speech recognition systems. Accuracy can be improved by adding a gender dependent acoustic model and unsupervised adaptation based on CMLLR (Constrained Maximum Likelihood Linear Regression). CMLLR-based batch-type unsupervised adaptation estimates a single global transformation matrix by utilizing prior unsupervised labeling, which unfortunately increases the computation time. Our proposed technique reduces prior gender selection and labeling time by using frame independent output probabilities of only gender dependent speech GMM (Gaussian Mixture Model) and context independent phoneme (monophone) HMM (Hidden Markov Model) in dual-gender acoustic models. The proposed technique further raises accuracy by employing a power term after adaptation. Simulations using spontaneous speech show that the proposed technique reduces computation time by 17.9 % and the relative error in correct rate by 13.7 % compared to the baseline without prior gender selection and unsupervised adaptation.

Index Terms: speech recognition, unsupervised adaptation, gender selection

1. Introduction

Users require not only highly accurate but also rapid speech recognition systems. Business demands higher accuracy with no increase in computation time compared to current baseline system; clients hesitate to increase the number of computers or to downsize the amount of speech data, so speed is often a key criterion to the introduction of new systems. Especially in the case of call centers, low-speed systems cannot handle the large amount of speech data stored each day. Our mission is to improve recognition accuracy and speed for stored speech with the same computer resources as used by the baseline system.

Our baseline speech recognition system adopted a conventional parallel decoding technique using dual-gender (male / female) acoustic models. Unfortunately, it reduces speed since its search space is larger than that of the 'ideal' single-gender dependent acoustic model. Imai *et al.* used context independent phoneme models (monophones) for 'online' speech detection and speech recognition with dual-gender models [1]. The monophone constraint for gender selection is also effective in reducing computation time in our 'offline' system. The proposed technique investigates the use of only gender dependent speech GMMs (Gaussian Mixture Models) and pause (silence) HMMs (Hidden Markov Models) in the gender selection process for a further reduction in computation time.

Accuracy depends, in part, on the speaker and the recording environment. Unsupervised adaptation based on MLLR (Maximum Likelihood Linear Regression) [2] is commonly used to improve accuracy. Adaptation is classified into two types: online and batch. The online type needs no prior labeling time since labels are derived from the recognition results of previous utterances; the negative side is that initial utterances receive no adaptation gain. The batch type can be expected to offer accuracy improvement for initial utterances. Initial utterances, in call-center speech situation such utterances often contain important details, are expected to be processed with high accuracy, so our proposed technique employs batch-type adaptation but the cost is that prior labeling is required.

Several adaptation techniques (e.g. [3]) that use a single global transformation matrix, like CMLLR (Constrained MLLR) [4], have been proposed. On the premise of estimating only a single matrix, we consider that the proposed technique can reduce computation time while realizing higher accuracy through rapid and appropriate labeling. One existing rapid unsupervised adaptation technique estimates the matrix based on coarse 2 class labels (speech / silence) using GMM [5]. Our proposal estimates the matrix based on fine (the 30 classes used in our system) labels using monophones [6]. The proposed technique increases the labeling time, but offers improved accuracy since it makes unsupervised adaptation from fine labeling, unlike the coarse labeling used in [5]. Moreover, the Forward-Backward algorithm [7] is conventionally used for statistics accumulation in calculating the matrix; the proposed technique realizes simple implementation and rapidity by using the approximate occurrence probabilities as determined from frame independent output probabilities. This approximation does not degrade accuracy in the experiments under the limited restrictions of gender is known and the amount of data is small [6]. Another advance is use of the power term after adaptation to improve accuracy.

This paper proposes a rapid combination technique of unsupervised adaptation and gender selection using the frame independent output probabilities with dual-gender acoustic models against a large amount of stored gender-unknown speech. Simulations show that the proposed technique offers rapidity and high accuracy; our technique reduces computation time and recognition errors significantly. The proposed technique is more practical because it does not require a change in the hardware configuration; it runs at over the baseline speed.

The rest of this paper is organized as follows; the proposed technique is described in Section 2. Section 3 introduces the experiments conducted that show the effectiveness of the proposed technique. Our conclusion is drawn in Section 4.

2. Proposed approach

Fig. 1 shows the framework of the proposed system. It consists of two parts; gender selection and unsupervised adaptation. The latter part has three components as follows; monophone con-

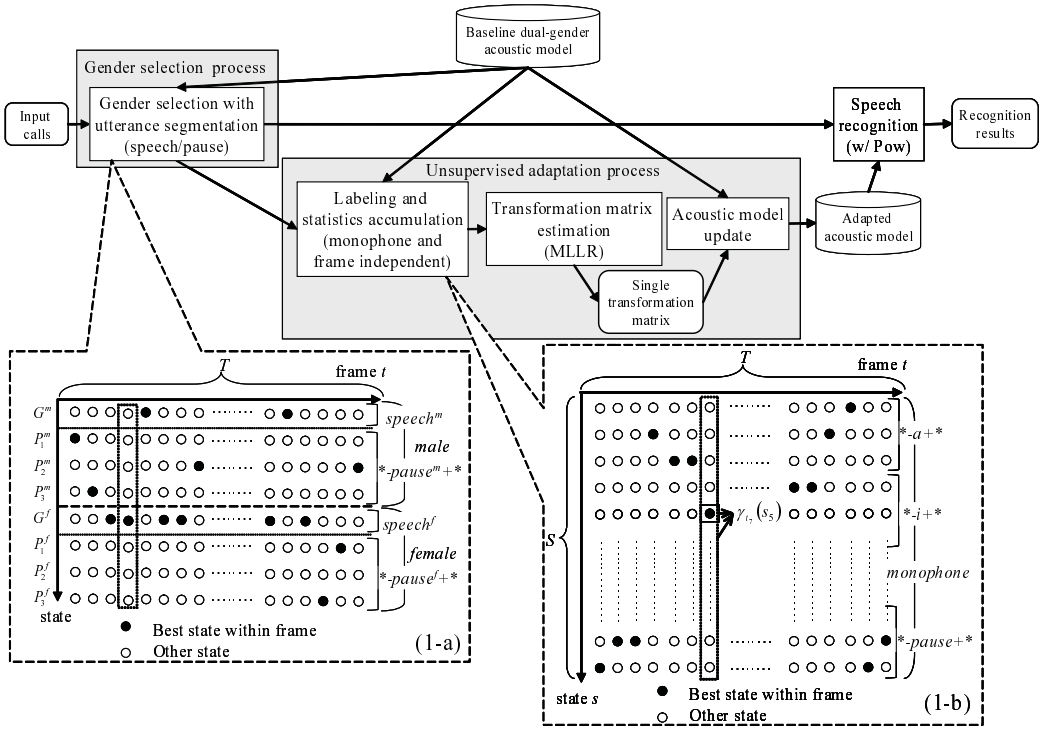

Figure 1: *Proposed system.*

straint, frame independent statistics accumulation, and power utilization. The proposed system performs gender selection utterance by utterance. It then applies fast frame independent statistics accumulation with monophone constraint against the utterances estimated to be from the same gender in adaptation process, and performs speech recognition with power utilization after adaptation.

2.1. Gender selection with utterance segmentation

2.1.1. Utterance segmentation

Utterance segmentation uses (gender dependent) speech GMMs (G^g) and pause HMMs (P_1^g, P_2^g, P_3^g; 3 states of the phoneme model used in our system) in dual gender ($g \in \{m, f\}$; m: male and f: female) dependent acoustic models. Utterance start-point is detected by a basic energy-based method with hangover time. After start-point detection, output probabilities $b_j(\mathbf{O}_t)$ of speech and pause models ($j \in G^g, P_1^g, P_2^g, P_3^g$) for feature vector $\mathbf{O}_t$ at frame t are calculated. If speech models ($b_{G^g}(\mathbf{O}_t)$) is the best state (● in Fig. 1-a), frame t is considered to be speech. If not, frame t is considered to be a pause. In the case that the pause frame is continued over τ^{pau} (e.g. 0.8 sec), the utterance is segmented as the end-point. Excessive utterance segmentation loses consonant discrimination and degrades accuracy in posterior speech recognition. Therefore, as the interval time between utterances is under τ^{intvl} (e.g. 1.0 sec), the utterances are concatenated. Whereas [1] uses monophones for segmentation, the proposed technique uses only the output probabilities of speech / pause model so it simplifies implementation and reduces the computation time.

2.1.2. Gender selection

The proposed technique selects gender concurrently with utterance segmentation. It determines gender by a majority vote of best state, either male ($b_{G^m}(\mathbf{O}_t)$) or female ($b_{G^f}(\mathbf{O}_t)$) within each speech frame. The above utterance concatenation is performed only between utterances of same gender. This gender selection only counts the number of best frames against each gender model, and so runs in less computation time.

2.2. Unsupervised adaptation

2.2.1. Monophone constraint

Lee *et al.* achieves rapid speech recognition by using monophones [8]. The proposed technique also speeds up unsupervised labeling by using only monophones; the assumption is that a monophone is an approximate model of a triphone. Our target is to estimate a single global transformation matrix. The single matrix obviously has fewer elements than the multiclass matrix. We consider that the sophisticated labeling provided by triphones is not required to estimate the fewer elements. Thus, it is sufficient to use monophones in labeling even if this yields a few errors. Triphones have many more states than monophones, so our state oriented monophone constraint can reduce computation time significantly.

2.2.2. Frame independent statistics accumulation

The transformation matrix is estimated using posterior probability $\gamma_t(s, m)$ from the m-th mixture component distribution of state s at frame t; $\gamma_t(s, m)$ is calculated from occurrence probability $\gamma_t(s)$ of state s. $\gamma_t(s)$ is usually estimated using the Forward-Backward algorithm [7], though occurrence probability $\gamma_t(s)$ is approximately calculated only from frame independent output probability $b_s(\mathbf{O}_t)$ of state s for feature vector $\mathbf{O}_t$ as shown in Eq. (1) to simplify implementation and reduce computation time. The unsupervised labeling process calculates $b_j(\mathbf{O}_t)$ of every state j in S states (● and ○ in Fig. 1-b) for $\mathbf{O}_t$ at each frame t, and only has to acquire $\sum_{j=1}^{S} b_j(\mathbf{O}_t)$ and $b_s(\mathbf{O}_t)$ of best state s (● in Fig. 1-b) within each frame t. S is the total number of states belonging to monophones as described in Section 2.2.1. Unlike the conventional prior labeling process, which needs to generate a number of recognition hypotheses to acquire the hypothesis with maximum likelihood using a large vocabulary word trigram, our frame independent statistics accumulation only has to calculate the output probabilities of the limited states, thus reducing the computation time. The well-known Viterbi training algorithm [9] approximates $\gamma_t(s)$ as be-

ing equal to 1.0 on the Viterbi path with the constraint of neighboring frames. Instead of this frame oriented constraint, our proposed technique uses state oriented adjustment to approximate $\gamma_t(s)$. Even if the labeling is not accurate in a frame, $\gamma_t(s)$ becomes smaller, so the influence of labeling error is reduced in our frame independent statistics accumulation.

$$\gamma_t(s) \simeq \begin{cases} \dfrac{b_s(\mathbf{O}_t)}{\sum_{j=1}^{S} b_j(\mathbf{O}_t)} & \text{if } s \text{ is best state at } t \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Posterior probability $\gamma_t(s,m)$ is calculated using the approximate occurrence probability $\gamma_t(s)$ shown in Eq. (2). Here, M_s is the number of distributions belonging to state s, $c_{s,m}$ is the m-th mixture weight and $N_{s,m}(\cdot)$ is the m-th multidimensional Gaussian distribution function with mean vector $\mu_{s,m}$ and covariance matrix $\Sigma_{s,m}$ of state s.

$$\gamma_t(s,m) \simeq \gamma_t(s) \cdot \frac{c_{s,m} N_{s,m}(\mathbf{O}_t|\mu_{s,m},\Sigma_{s,m})}{\sum_{k=1}^{M_s} c_{s,k} N_{s,k}(\mathbf{O}_t|\mu_{s,k},\Sigma_{s,k})} \quad (2)$$

The statistics of mean parameter, $\sum_{t=1}^{T} \gamma_t(s,m) \cdot \mathbf{O}_t$ and $\sum_{t=1}^{T} \gamma_t(s,m)$, are accumulated using posterior probability $\gamma_t(s,m)$ over the total number of frames, T. The single global transformation matrix is generated from these accumulated statistics using the model-space MLLR of [4]. The mean parameters of all distributions in the acoustic model (triphones as well as monophones) are transformed by this matrix.

2.2.3. Power utilization

Speech power changes due to the speaker and microphone position. The power term must be utilized, by applying a power adapted model, if the speech recognition accuracy is to be raised significantly. The proposed technique uses the power term only after adaptation, not before; occurrence probability $\gamma_t(s)$ is calculated using the likelihood without power while $N_{s,m}(\cdot)$ is calculated with power in Eq. (2). Furthermore, speech power level is normalized utterance by utterance in acoustic model training.

3. Experiments

3.1. Experimental conditions and task

Table 1 shows the speech analysis conditions, Table 2 shows the acoustic model (HMM) parameters used in the experiments, and Table 3 shows the evaluation task.

Table 4 shows the techniques compared. The "1. baseline" is the technique using parallel decoding without prior gender selection and unsupervised adaptation.

At first, the influence of gender selection is investigated. The effect of gender selection, "GS", is confirmed by comparing gender-known, "gd", to -unknown "m/f"; "gd" uses the 'ideal' gender dependent acoustic model. The proposed gender selection with speech / pause models, "GS(s/p)", is compared to monophone-based technique, "GS(mono)", like that of [1]. Then, the proposed unsupervised adaptation technique, "pUA", using monophone constraint labeling and frame independent statistics accumulation, is then compared to the conventional unsupervised adaptation technique, "cUA", using the Forward-Backward statistics accumulation with 'ideal' gender dependent acoustic model and labeling by several language models; speech / pause loop "cUA(s/p)" like [5], monophone loop "cUA(mono)" and word trigram "cUA(tri)". Finally, the effect of power utilization (w/Pow: with power) is verified.

Table 1: *Speech analysis conditions.*

Sampling rate	16 [kHz]
Window type	Hamming
Frame width	20 [msec]
Frame shift	10 [msec]
Feature parameters	MFCC 12, ΔMFCC 12, Δpower, or partial power after adaptation

Table 2: *Acoustic model. (male / female)*

Conditions	
HMM	Continuous mixture distribution
Gender	Dual gender dependent
Phoneme model	Triphone and monophone
# of parameters	
States in monophones	90
States	1,958
Distributions	26,568 / 29,836
Phonemes	30
Training data	
Size	122.71 / 113.23 [hour]
# of utterances	109294 / 110792

Table 3: *Evaluation task.*

Utterance style	Spontaneous dialog
Size	9.91 [hour]
# of calls	120
# of utterances	9,056
Speakers	7 males / 17 females
Language	Japanese
Language model	Word trigram
Vocabulary size	59,676 words
Speech recognition decoder	VoiceRex [10]

3.2. Experimental results

The evaluation results are shown in Table 5 and Table 6. We utilized a character-unit evaluation to eliminate the influence of word length; the abbreviations of "Cor." and "Acc." mean correct rate and accuracy, respectively. The computation time is normalized by the "1. baseline" recognition time; "Slct." is the gender selection time, "Adpt." is the adaptation time, and "Sum." is the total computation time.

The effect of gender selection is shown in Table 5. "1. baseline" degraded speed and accuracy compared to 'ideal' gender dependent technique, "2. gd", since it expanded the search space and generating gender selection error. The prior gender selection techniques (3 and 4) achieved equivalent accuracy to the ideal technique, "2. gd", and the proposed "4. m/f+GS(s/p)" is slightly faster than "3. m/f+GS(mono)" with equivalent accuracy.

The effect of unsupervised adaptation is shown in the upper part of Table 6 without power utilization. All unsupervised techniques (5-8) offered better accuracy compared to the no adaptation techniques (2 or 4). The conventional unsupervised adaptation technique using word trigram achieved the best accuracy but its computation time is twice of "1. baseline" and it also requires the 'ideal' gender dependent acoustic model. Simplifying the language model in labeling (5→7), reduced not only computation time but also accuracy. The conventional speech / pause based technique, "7. gd+cUA(s/p)", similar to [5], is faster than "1. baseline", but it offers the

Table 4: *Compared Techniques.*

ID and name	Gender selection	Labeling and statistic accumulation	Power utilization
1. m/f: baseline	Parallel decoding	None	False
2. gd	Known	None	False
3. m/f+GS(mono)	Monophone loop	None	False
4. m/f+GS(s/p)	**Speech / pause loop**	**None**	**False**
5. gd+cUA(tri)	Known	Word trigram and Forward-Backward	False
6. gd+cUA(mono)	Known	Monophone loop and Forward-Backward	False
7. gd+cUA(s/p)	Known	Speech / pause loop and Forward-Backward	False
8. 3+pUA(mono)	**Speech / pause loop**	**Monophone's state loop and frame independent**	**False**
9. 5+w/Pow	Known	Word trigram and Forward-Backward	True
10. 6+w/Pow	Known	Monophone loop and Forward-Backward	True
11. 7+w/Pow	Known	Speech / pause loop and Forward-Backward	True
12. 8+w/Pow: proposed	**Speech / pause loop**	**Monophone's state loop and frame independent**	**True**

Table 5: *Performance of the proposed prior gender selection.*

ID and name	Cor.	Acc.	Sum.	Slct.
1. m/f: baseline	79.22	73.79	1.00	-
2. gd	80.78	75.39	.939	-
3. m/f+GS(mono)	80.72	75.40	.977	.0463
4. m/f+GS(s/p)	**80.71**	**75.34**	**.956**	**.00775**

Table 6: *Performance of the proposed unsupervised adaptation.*

ID and name	Cor.	Acc.	Sum.	Slct.	Adpt.
5. gd+cUA(tri)	82.04	76.76	2.00	-	1.08
6. gd+cUA(mono)	81.61	76.26	1.46	-	.512
7. gd+cUA(s/p)	81.32	76.02	.936	-	.0217
8. 4+pUA(mono)	**81.63**	**76.38**	**.922**	**.00777**	**.0195**
9. 5+w/Pow	82.07	77.15	1.92	-	1.08
10. 6+w/Pow	81.62	76.58	1.39	-	.512
11. 7+w/Pow	80.01	74.80	.902	-	.0222
12. 8+w/Pow	**82.07**	**77.00**	**.820**	**.00775**	**.0201**

least improvement in accuracy. The proposed unsupervised adaptation technique, "8. 4+pUA(mono)", provided the equivalent accuracy to the conventional monophone-based technique, "6. gd+cUA(mono)", and the total computation time was also under that of "1. baseline"; the approximation of Eq. (1) is valid and the proposed adaptation process contributed to higher speed due to the improving beam search efficiency provided by the adaptation effect.

With power utilization, see the lower part of Table 6, most techniques demonstrate some gain. As a result, the proposed technique, "12. 8+wPow", achieved significant improvements in both accuracy and rapidity. Moreover, it offers over twice the speed of the conventional trigram-based adaptation technique with little degradation in accuracy.

4. Conclusions

This paper proposed rapid unsupervised batch adaptation using frame independent output probabilities of speech (gender) / pause / monophone models. The proposed technique segments utterances and selects gender simultaneously with less computation time than is possible if gender and pause models are used. It offers rapid unsupervised adaptation using monophone constraint and frame independent statistics accumulation against selected gender dependent acoustic model. After the adaptation, it uses a power term in speech recognition process to improve accuracy. Tests showed that our technique reduced the relative error in correct rate by 13.7 % and computation time by 17.9 % compared to the baseline without prior gender selection and unsupervised adaptation.

5. Acknowledgements

We are grateful to our project manager, Mr. Haruhiko KOJIMA of NTT Cyber Space Laboratories, for giving us the opportunity to pursue this work. We also wish to thank the members of the Speech, Acoustics and Language Laboratory for their helpful advice.

6. References

[1] T. Imai, S. Sato, A. Kobayashi, K. Onoe, and S. Homma, "Online speech detection and dual-gender speech recognition for captioning broadcast news," in *Proc. Interspeech 2006 - ICSLP*, Pittsburgh, Sep. 2006, pp. 1602–1605.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] S.-A. Selouani and D. O'Shaughnessy, "Speaker adaptation using evolutionary-based linear transform," in *Proc. Interspeech 2006 - ICSLP*, Pittsburgh, Sep. 2006, pp. 1109–1112.

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

[5] S. S. Kozat, K. Visweswariah, and R. Gopinath, "Efficient, low latency adaptation for speech recognition," in *Proc. ICASSP 2007*, vol. 4, Honolulu, Apr. 2007, pp. 777–780.

[6] S. Kobashikawa, A. Ogawa, Y. Yamaguchi, and S. Takahashi, "Rapid unsupervised adaptation using context independent phoneme model," in *Proc. IEEE International Symposium on Consumer Electronics*, Kyoto, May 2009, pp. 209–212.

[7] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. IEEE*, vol. 77, no. 2, Feb. 1989, pp. 257–286.

[8] A. Lee, T. Kawahara, and K. Shikano, "Gaussian mixture selection using context-independent HMM," in *Proc. ICASSP 2001*, vol. 1, Salt Lake City, May 2001, pp. 69–72.

[9] B.-H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden markov models," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 38, no. 9, pp. 1639–1641, 1990.

[10] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "VoiceRex – spontaneous speech recognition technology for contact-center conversations," *NTT Tech. Rev.*, vol. 5, no. 1, pp. 22–27, 2007.