Mispronunciation Detection and Diagnosis in L2 English Speech Using Multi-Distribution Deep Neural Networks

Kun Li and Helen Meng

Human-Computer Communications Laboratory
Department of System Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong SAR, China
{kli, hmmeng}@se.cuhk.edu.hk

Abstract

This paper investigates the use of multi-distribution Deep Neural Networks (DNNs) for mispronunciation detection and diagnosis (MD&D). Our existing approach uses extended recognition networks (ERNs) to constrain the recognition paths to the canonical pronunciation of the target words and the likely phonetic mispronunciations. Although this approach is viable, it has some problems: (1) deriving appropriate phonological rules to generate the ERNs remains a challenging task; (2) the acoustic model (AM) and the phonological rules are trained independently and hence contextual information is lost; and (3) phones missing from the ERNs cannot be recognized even if we have a well-trained AM. Hence we propose an Acoustic Phonological Model (APM) using a multi-distribution DNN, whose input features include acoustic features and corresponding canonical pronunciations. The APM can implicitly learn the phonological rules from the canonical productions and annotated mispronunciations in the training data. Furthermore, the APM can also capture the relationships between the phonological rules and related acoustic features. As we do not restrict any pathways as in the ERNs, all phones can be recognized if we have a perfect APM. Experiments show that our method achieves an accuracy of 83.3% and a correctness of 88.5%. It significantly outperforms the approach of forced-alignment with ERNs whose correctness is 75.9%.

Index Terms: speech recognition, mispronunciation detection and diagnosis, L2 English speech, deep neural networks

1. Introduction

Automatic mispronunciation detection and diagnosis (MD&D) is the core of Computer-Aided Pronunciation Training (CAPT) systems. It can be treated as a special type of automatic phone recognition. When the recognized phones differ from the canonical productions (obtained from the text prompts presented to the speaker), mispronunciation detection and diagnosis are achieved respectively.

Generally speaking, there are two traditional ways to do MD&D. One is that we can simply treat it as a free phone recognition task and apply the automatic speech recognition (ASR) technology to it. State-of-the-art ASR systems use Hidden Markov Models (HMMs) to model the sequential structure of speech signals [1]. Traditionally, Gaussian Mixture Models (GMMs) are used to model the conditional distribution of speech signal spectrum for each HMM state. Recently, due to the development of highly effective learning techniques like Deep Neural Networks (DNNs) [2, 3], DNNs are used to replace GMMs as part of acoustic models and achieved significant improvements [4, 5, 6]. Many derivative types of DNNs, such as Convolutional Neural Networks (CNNs) [7, 8] and Recurrent Neural Networks (RNNs) [9], also achieved impressive improvements. Their phone recognition error (PRE) rates over the TIMIT corpus are below 20% [6, 8, 9]. However, due to the deviations of second language (L2) speech from native productions, this method tends to achieve poor performance even if we adapt the acoustic models (AMs) with L2 English speech. One reason is that we do not have sufficient L2 English speech to cover all the deviations in the L2 acoustic space.

Another way is that we align the phones using extended recognition networks (ERNs) which cover not only the canonical pronunciation of words, but also the likely phonetic mispronunciations [10, 11, 12, 13, 14]. Performing phonetic recognition with such ERNs involve identifying the path with the highest probability, which is taken as the phonetic transcription of the uttered word(s) by the learner. To achieve better performance, we should improve the AMs by using DNNs to replace GMMs [14], or improve the ERNs to incorporate as many as possible phonological rules with high precision and recall rates [10, 11, 12]. This method generally achieves better performance than the free phone recognition. However, there are still some problems: (1) it is difficult to build ERNs that incorporate as many as possible mispronunciation paths with high precision and recall rates. In [10], a set of 51 phonological rules are carefully designed by an expert, and the precision and recall rates are 14% and 46%, respectively; while in [11], a set of 216 rules are selected by a data-driven method from 100 native Cantonese speakers, whose precision and recall rates are 31% and 63%, respectively. Poor performances are due to misreading the text prompts by subjects and guessed pronunciations for words unfamiliar to the speakers [11, 13]. (2) Forced-alignment with ERNs trains the AMs and phonological rules independently; hence contextual information is lost. (3) Phones missing from the ERNs cannot be recognized even if we have a well-trained AM.

To overcome the above drawbacks, we propose an Acoustic Phonological Model (APM) which uses a multi-distribution DNN to incorporate acoustic features and phonological rules. In the CAPT system, the prompts for L2 learners to utter are carefully designed, thus we can extract the canonical pronunciations. During transcription, the annotators are also aware of the canonical pronunciations. Based on these, we try to train a DNN using acoustic features and corresponding canonical pronunciations to infer the actual pronunciations of L2 learners that match the annotation results with the highest probability. We believe that this DNN can automatically learn the phonological rules from the canonical pronunciations and annotation results, and can further mine the relationship between acoustic features and phonological rules.

As the input features of APM include acoustic features (which are assumed to have Gaussian distribution) and

corresponding canonical pronunciations (which can be set as binary), a multi-distribution DNN is used in this work. Multi-distribution DNNs have been applied to speech synthesis [15] and lexical stress detection [16]. Similar to traditional DNNs, they are also constructed by stacking up multiple Restricted Boltzmann Machines (RBMs) on top of one another. Excluding the bottom RBM, all the other ones are traditional Bernoulli RBM (B-RBM), whose hidden and visible units are all binary. The bottom RBM is a type of mixed Gaussian-Bernoulli RBM (GB-RBM), whose hidden units are binary while some visible units are Gaussian distributed and the other visible units are binary.

In this work, we first realize traditional free phone recognition. A monophone AM and a 5-gram phone-based language model (LM) are built, both of which use DNNs. Then an APM incorporating acoustic features and phonological rules is built. The structure of our paper is designed as follows: Section 2 describes the free phone recognition for L2 English; Section 3 introduces our approach using the APM; Sections 4 and 5 present the experimental results and conclusions, respectively.

2. Free phone recognition for L2 English

To realize free phone recognition for L2 English, an AM and a LM are built, both of which use DNNs.

2.1. Acoustic model (AM)

To reduce complexity, we only realize a monophone AM, whose structure is shown in Fig. 1.

2.1.1. MFCC features

The speech is sampled at 16 kHz. To compensate for the high-frequency part of speech signal, a pre-emphasis filter is applied to the speech, whose transform function is $1-0.97z^{-1}$. Then Fast Fourier Transform is performed on a 25-ms Hamming window with a 10-ms frame shift. Finally, a set of 13 Mel-frequency cepstral coefficient (MFCC) features are computed for every 25-ms frame. These features are normalized to have zero mean and unit variance.

2.1.2. Architecture of DNN

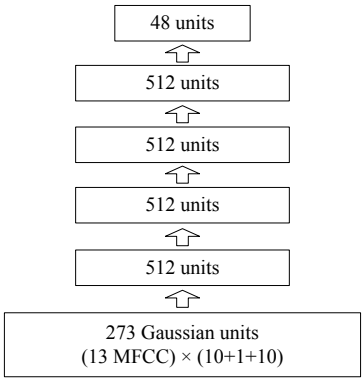

Figure 1: *Structure of the monophone acoustic model.*

In our experiments, we use 21 frames (10+1+10) of MFCCs as the input features of the DNN, thus there are 273 Gaussian units in the bottom of the DNN. Besides the bottom layer, there are 4 hidden layers and one top layer. There are only 48 units in the top output layer, for we use a 48-phone set [1] and it is a monophone AM, i.e. each phone has only one state. Following [1], this 48-phone set will be mapped to a 39-phone set in the final step of decoding.

2.2. Language Model (LM)

As we are using a monophone AM, we can simply build a LM to generate the probabilities of phone (state) transitions. In [1], it shows that a phone LM helps improve the performance of phone recognition. In this work, we use a DNN to construct a 5-gram phone-based LM.

The structure of the DNN is shown in Fig. 2. There are 195 binary input units indicating the presence or absence of the corresponding phone. The top layer is a "softmax" layer with 39 units (phones). Note that it is different with the top layer of the DNN in Fig. 1, which is generated from a traditional sigmoid function instead of a softmax function. We treat these two top layers differently, because it is convenient to use the softmax function to "normalize" the probabilities so that the output probabilities of the LM sum to one. However, for the DNN in Fig. 1, we use a 48-phone set, in which two or more phones maybe mapped to the same phone in a 39-phone set, e.g., the phones /ix/ and /ih/ are mapped to /ih/ in the 39-phone set.

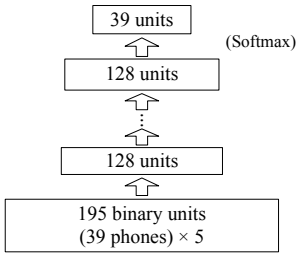

Figure 2: *Architecture of the 5-gram phone-based language model.*

2.3. Phone Recognition

In Viterbi decoding, the phone (state) sequence with the largest conditional probability is determined as the recognized phone sequence, as given in Eq. (1):

$$\hat{q} = arg\max_q p(\boldsymbol{q} \mid \boldsymbol{X}) \quad (1)$$

where $\boldsymbol{X}$ is the sequence of acoustic feature vectors, $\boldsymbol{q}$ denotes a phone sequence.

The probability of $\boldsymbol{q}$ given $\boldsymbol{X}$ is:

$$p(\boldsymbol{q}\,|\boldsymbol{X}) = p(q_1 \mid \boldsymbol{X})p(q_2 \mid q_1, \boldsymbol{X}) \cdots p(q_t \mid q_1, \cdots, q_{t-1}, \boldsymbol{X}) \cdots$$
$$\approx p(q_1|\, x_1)p(q_2|q_1, x_2) \cdots p(q_t|q_{t-5}, \cdots, q_{t-1}, x_t) \cdots \quad (2)$$

where x_t is the acoustic feature vector of the t^{th} frame, q_t denotes the phone at the t^{th} frame. Note that we use a 5-gram LM and x_t has a context windows of (10 + 1 + 10) frames.

Applying Bayes' Theorem, we have:

$$p(q_t|q_{t-5}, \cdots, q_{t-1}\, x_t) = \frac{p(q_t)p(q_{t-5}, \cdots, q_{t-1},\, x_t|q_t)}{p(q_{t-5}, \cdots, q_{t-1},\, x_t)}$$
$$\approx \frac{p(q_t)p(q_{t-5}, \cdots, q_{t-1}|q_t)p(\, x_t|q_t)}{p(q_{t-5}, \cdots, q_{t-1})p(x_t)}$$
$$= p(q_t|q_{t-5}, \cdots, q_{t-1})\frac{p(q_t|x_t)}{p(q_t)} \quad (3)$$

From Eq. (2) and (3), we have:

$$p(\boldsymbol{q} \mid \boldsymbol{X}) \approx p(q_1 \mid x_1)p(q_2 \mid q_1)\frac{p(q_2 \mid x_2)}{p(q_2)}\cdots$$
$$p(q_t \mid q_{t-5},\cdots,q_{t-1})\frac{p(q_t \mid x_t)}{p(q_t)}\cdots \quad (4)$$

where $p(q_t)$ is the prior probability and $p(q_t \mid x_t)$ is the posterior probability which can be calculated from the DNN of AM, $p(q_t \mid q_{t-5},\cdots,q_{t-1})$ is the phone transition probability that can be computed from the 5-gram phone-based LM.

3. Acoustic Phonological Models (APMs)

In this section, we propose an APM which uses a multi-distribution DNN to incorporate acoustic features and phonological rules. We first align the canonical pronunciation with L2 English speech, and then use the APM to calculate the posterior probabilities of each phone. In the final step we introduce the Viterbi decoding which is similar to the last section.

3.1. Forced-alignment of canonical pronunciation

In order to get each frame's corresponding expected phone and the context phones, i.e., their preceding and following canonical phones, we use the AM trained in Section 2.1 to align the speech with their canonical pronunciation, which can be derived from dictionary according to the words prompted to readers.

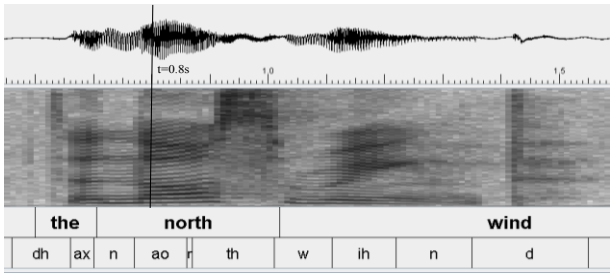

Figure 3: *An example of L2 English speech aligned with canonical phones.*

3.2. Structure of APM

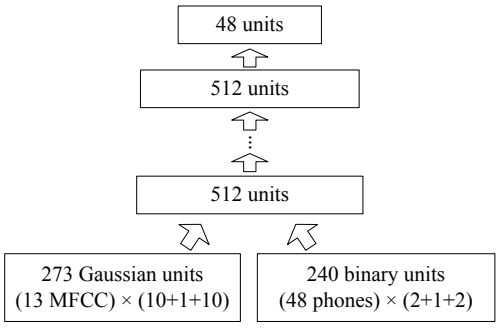

Figure 4: *Structure of the Acoustic Phonological Model.*

Similar to the AM in Section 2.1, we use 273 MFCC features which are assumed to have Gaussian distribution. From the forced-alignment of canonical pronunciation, we have each frame's expected canonical phone and their context phones. For the example in Fig. 3, the expected phone for frame t = 0.8s is /ao/, and its preceding and following phones are /n/ and /r/, respectively.

The structure of our APM is shown in Fig. 4, which is a multi-distribution DNN [15, 16]. There are 273 Gaussian and 240 binary visible units in the bottom of the DNN. The other layers are similar to those in Fig. 1.

3.3. Phone Recognition

Using the APM, the object function in Viterbi decoding is changed from Eq. (1) to Eq. (5):

$$\hat{q} = arg\max_q p(\boldsymbol{q} \mid \boldsymbol{X}, \boldsymbol{q^{Dict}}) \quad (5)$$

where $\boldsymbol{q^{Dict}}$ is the canonical pronunciation.

Similar to Eq. (4), we can have:

$$p(\boldsymbol{q} \mid \boldsymbol{X}) \approx p(q_1 \mid x_1, q_1^{Dict})p(q_2 \mid q_1)\frac{p(q_2 \mid x_2, q_2^{Dict})}{p(q_2)}\cdots$$
$$p(q_t \mid q_{t-5},\cdots,q_{t-1})\frac{p(q_t \mid x_t, q_t^{Dict})}{p(q_t)}\cdots \quad (6)$$

where q_t^{Dict} is the corresponding canonical pronunciation with a context window of (2+1+2) phones at the t^{th} frame. Note that $p(q_t \mid x_t, q_t^{Dict})$ can be computed from the APM.

4. Experiments

4.1. Corpus

Our experiments are based on the TIMIT and CHLOE (**Ch**inese **L**earners **o**f **E**nglish) corpus. For the CHLOE corpus, only Mandarin learners' data are used, which contains 110 speakers (60 males and 50 females). There are five parts in CHLOE: confusable words, minimal pairs, phonemic sentences, the Aesop's Fable "The North Wind and the Sun" and prompts from TIMIT. Excluding the TIMIT prompts, all the other parts are labeled by trained linguists.

To transcribe the L2 English speech of CHLOE, we first built acoustic models using HTK [18] based on the TIMIT corpus to align the canonical pronunciations with the L2 English speech. Then our linguists annotated the speech with actual pronunciations. To save labor, our linguists mainly focused in labeling (modifying) the phone sequences, thus the accuracy of the phone boundaries is not high. Hence, these annotated phone sequences should be re-aligned using the AMs described in Section 2. We do the forced-alignment and train the AMs iteratively until there is no significant improvement.

We randomly split CHLOE by speakers into a training set which contains 88 speakers and a test set containing 22 speakers (12 males and 10 females). For this training set, only about 30% of the data are labeled. The details of the TIMIT and CHLOE corpus are shown in Table 1. Note that the data of the TIMIT corpus are native English speech and are also used as part of our training data.

Table 1. *Details of corpus used in our experiments.*

	TIMIT	CHLOE	
	Train	Train	Test
Speakers	630	88	22
Unlabeled	---	40 h	---
Labeled	4.5 h	15.5 h	4 h

4.2. DNNs training

The DNNs training in this work is similar to [14, 16]. In the pre-training stage, we try to maximize the log-likelihood of RBMs. One-step Contrastive Divergence (CD) [2] is used to approximate the stochastic gradient. 20 epochs are performed with a batch size of 256 frames. For the parameters of GB-RBM, a learning rate of $\eta = 0.002$ is used; while for the parameters of B-RBMs, a learning rate of $\eta = 0.005$ is used. Increment in each batch is smoothed by a momentum of $\gamma = 0.9$, thus we have the following update rule for the t^{th} increment of θ: $\Delta\theta^{(t+1)} = \Upsilon\Delta\theta^{(t)} + \eta\frac{\partial\mathcal{L}}{\partial\theta}$, where $\frac{\partial\mathcal{L}}{\partial\theta}$ is the gradient. In the fine-tuning stage, the standard back-propagation algorithm [17] is performed.

4.3. Results of phone recognition

The experimental results of phone recognition are shown in Table 2. It shows that the correctness of the free phone recognition using the traditional AM in Section 2.1 is only 74.0%. In [11], a method of forced-alignment using ERNs achieved a correctness of 75.9%, although it used GMMs as its acoustic models. It shows that ERNs generated by appropriate phonological rules improve the performance of L2 English phone recognition.

Our new approach using the APM outperforms the above two methods and achieves significant improvements. Its correctness and accuracy are 88.5% and 83.3%, respectively. The difference between the traditional AM and the APM is that the later one makes use of the canonical pronunciations. From the canonical pronunciations and annotated results, the APM can automatically learn the phonological rules. Furthermore, the APM can also capture the relationship between the acoustic features and phonological rules. Using the canonical pronunciations, the APM can slightly outperform the complicated models for free phone recognition of native English speech, the best accuracy of which as far as we know is about 82.3% [5, 8, 9].

If we increase the hidden units in each layer of the DNN from 256 to 512, the accuracy is improved from 82.1% to 83.3%. Regardless of the limitation of computing power, furthermore improvement maybe gained if more units of each hidden layer are used.

Table 2. *Performance of phone recognition with different methods.*

	Nodes #	Correct	Acc.
ERNs (GMMs) [11]	---	75.87%	---
Traditional AM	256	73.96%	57.50%
APM	256	87.89%	82.08%
	512	**88.52%**	**83.25%**

Note: The starting and ending silences are not counted here. The correctness and accuracy are calculated following [18]: $Correctness = \frac{N-S-D}{N}$, $Accuracy = \frac{N-S-D-I}{N}$; where N is the total number of labels; while S, D and I denote for the counts of substitution, deletion and insertion errors, respectively.

4.4. Preliminary results of MD&D

As introduced in Section 1, when the recognized phones differ from the canonical productions, mispronunciation detection and diagnosis (MD&D) are achieved respectively. Thus the phone accuracy is one of the most important metrics to evaluate the performance of MD&D. In this sub section, we also provide more but preliminary experimental results of MD&D for comparing with other approaches.

Table 3 shows the results of MD&D with different approaches. The AM for free phone recognition achieves poor performance, whose false rejection is 31.0%. Large false rejection is due to many insertion and substitution errors. In [13], the ERNs are generated not only from "phoneme-to-mispronunciation conversion" but also from "grapheme-to-mispronunciation conversion", and it achieves a false rejection of only 25.6%. The false rejection of the APM with 512 nodes in each hidden layer is only about 9.3%.

Table 3. *Performance of mispronunciation detection and diagnosis with different approaches.*

	canonicals		mispronunciations		
	True Accept.	False Rejection	False Accept.	True Rejection.	
				Correct Diag.	Diag. Error
ERNs [13]	74.37%	25.63%	22.80%	54.55%	45.45%
AM (256)	69.05% (27,548)	30.95% (12,346)	**18.79%** (1,143)	67.64% (3,342)	32.36% (1,599)
APM (256)	**91.21%** (36,575)	**8.79%** (3,524)	44.92% (2,771)	76.43% (2,597)	23.57% (801)
APM (512)	90.72% (36,389)	9.28% (3,721)	39.64% (2,452)	**80.77%** (3,016)	**20.34%** (718)

Note: Only about 11% of the phones are mispronounced by L2 learners.

5. Conclusions

In this paper, we investigate mispronunciation detection and diagnosis (MD&D) using multi-distribution DNNs. We first build a traditional acoustic model (AM) using a DNN to align the canonical pronunciations with L2 English speech. Then we construct an Acoustic Phonological Model (APM) also using a DNN, whose input features include MFCC features which are assumed to have Gaussian distribution, as well as the corresponding canonical pronunciations which are binary values. The APM can implicitly learn the phonological rules from the canonical productions and annotated mispronunciations in the training data. Furthermore, the APM can capture the relationship between phonological rules and related acoustic features. In the final step of Viterbi decoding, a 5-gram phone-based language model is built with a DNN. Comparing with the forced-alignment with extended recognition networks (ERNs), which constrains the recognition paths to some most possible pronunciations and cannot recognize the phones missing from the ERNs, our method is simpler and more effective. Experimental results show that the free phone recognition using a traditional AM for L2 English speech achieves poor performance with a correctness of 74.0% and an accuracy of 57.5%. The force-alignment with ERNs using GMMs as AMs can achieve a correctness of 75.9%. Our method using an APM can gain a significant improvement, whose correctness and accuracy are 88.5% and 83.3%, respectively.

6. Acknowledgements

The work is partially supported by a grant from the HKSAR Government GRF (project number 415511).

7. References

[1] Lee, K-F and Hon, H-W, “Speaker-independent phone recognition using hidden Markov models”, IEEE Trans. on Audio, Speech and Language Proc., 1989.

[2] Hinton, G. E., Osindero, S. and Teh, Y., “A fast learning algorithm for deep belief nets”, Neural Computation, 18:1527-1554, 2006.

[3] Hinton, G. E. and Salakhutdinov, R. R., “Reducing the Dimensionality of Data with Neural Networks”, Science, 313:504-507, 2006.

[4] Mohamed, A., Dahl, G. E. and Hinton G. E., “Acoustic modeling using deep belief networks”, IEEE Trans. on Audio, Speech and Language Proc., 20(1):14-22, 2012.

[5] Dahl, G. E., Yu, D., Deng, L. and Acero A., “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition”, IEEE Trans. on Audio, Speech and Language Proc., 20(1):30-40, 2012.

[6] Mohamed, A., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E. and Picheny, M.A.“Deep belief networks using discriminative features for phone recognition”, ICASSP 2011.

[7] Abdel-Hamid, O., Mohamed, A., Jiang, H. and Penn, G., “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition”, ICASSP 2012.

[8] Deng L., Abdel-Hamid, O. and Yu D., “A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion”, ICASSP 2013.

[9] Graves, A., Mohamed A. and Hinton G. E., “Speech Recognition with Deep Recurrent Neural Networks”, ICASSP 2013.

[10] Harrison, A. M., Lo, W., Qian, X. and Meng, H., "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training", SLaTE 2009.

[11] Lo W., Zhang S. and Meng H., "Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System," Interspeech 2010.

[12] Qian X., Soong F. and Meng, H., "Discriminative Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer-Aided Pronunciation Training (CAPT)," Interspeech 2010.

[13] Qian, X., Meng, H. and Soong, F., "On Mispronunciation Lexicon Generation using Joint-sequence Multigrams in Computer-Aided Pronunciation Training (CAPT)", Interspeech 2011.

[14] Qian, X., Meng, H. and Soong, F., "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training", Interspeech 2012.

[15] Kang, S., Qian, X. and Meng, H., “Multi-distribution deep belief network for speech synthesis”, ICASSP 2013.

[16] Li, K., Qian, X., Kang, S. and Meng, H., “Lexical Stress Detection for L2 English Speech Using Deep Belief Networks”, Interspeech 2013.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors”, Nature, vol. 323, no. 6088, pp.533–536, 1986

[18] Yong, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., “The HTK book (for HTK version 3.4)”, Cambridge university, 187-189, 2006.