A Language Model-Based Design of Reduced Phoneme Set for Acoustic Model

Shuji Komeiji* and Toshihisa Tanaka†
* Tokyo University of Agriculture and Technology, Japan
E-mail: komeiji@sip.tuat.ac.jp Tel/Fax: +81-42-388-7123
† Tokyo University of Agriculture and Technology, Japan
E-mail: tanakat@cc.tuat.ac.jp Tel/Fax: +81-42-388-7123

***Abstract*—A language model-based design of reduced phoneme set for acoustic model is proposed. In the case where the amount of training data is too small to train each phoneme model, the reduction of the phoneme set can lead to a reduced discriminative model of phonemes, which can increase homophones that yield degradation of speech recognition. The proposed approach enables us to reduce phonemes preventing the degradation, regarding pronunciation/word sequence confusion rate calculated from n-grams in a language model. In an experiment, the phoneme set designed with proposed approach was applied to Japanese large vocabulary speech recognition system. The word error rate with full 39 phonemes set was 9.5%, while the error rate with the 10 phonemes set designed with the proposed approach was 11.1%. The degradation was able to be prevented within 2%.**

I. Introduction

An acoustic model (AM) which represents the statistical properties of speech is used in various fields such as automatic speech recognition (ASR), speaker recognition, speaker verification, language recognition, speech synthesis, etc [1], [2]. Recently, this technology has also been applied to the field of brain machine interface (BMI) for the possibility of decoding the brain activity during speaking, listening or imagining [3], [4], [5], [6], [7].

Focusing on the training of an AM, it is the important issue to determine the number of parameters to be trained within the amount of given training data. In typical ASR systems, a sequence of three phonemes named as triphone is likely to be used for an AM. Generally speaking, the total number of triphones is so large that not all of the triphones can be trained. For example, if the size of the phoneme set is 40, the total number of triphones amounts to 64,000 and the appearance frequencies of triphones vary widely [8]. The triphones that cannot be trained by sufficient data can degrade the accuracy of the ASR system. An approach to overcoming this issue is a decision tree clustering [9] that reduces the total number of triphones. The decision tree clustering is the de facto standard approach in present ASR systems.

A tiny amount of data ends up to being harmful in a speaker adaptation for an ASR, which is a technique for additional training of an AM with small amount of training data (called adaptation data) of specific speaker. The speaker adaptation can be done properly against various amount of adaptation data by autonomous model complexity control (AMCC) [10] to take advantage of a tree structured Gaussian mixture model (GMM) [11] as an AM. In the tree structured GMM, each leaf node corresponds to an original Gaussian distribution. These parent nodes correspond to the Gaussian distributions whose parameters share these leaf nodes parameters. The parameters of the root node corresponds to the common characteristic of whole phonemes. The AMCC controls the amount of parameters to be adapted by determining the depth of the tree structure based on the amount of adaptation data.

Another example of reduced parameters in an AM is a speech BMI. Herff et. al. applied the framework of an ASR decoder to the decoding of the speech related brain activity from the intracranial electrocorticogram (ECoG) [3]. In the decoding, a statistical model corresponding to an AM is trained from ECoG data. Due to the very limited amount of data to train the model, they reduced the size of the basic phoneme set by merging similar phonemes together. However, the reduction of the basic phoneme set for the control of the number of parameter is harmful for the decoding accuracy because of the increase of homophone words. Conventional works take account of not the increase of homophone words but only the acoustic similarity.

Therefore, in this paper, a design of reduced phoneme set based on a language model (LM) is proposed. To introduce a LM for the reduction, pronunciation/word sequence confusion rate (PWCR) is defined to estimate a degradation of an accuracy of ASR with the reduced phoneme set.

In Section II, the related works including ASR and reduction of phoneme set are described. In Section III, the PWCR calculated by a LM is defined to evaluate the degradation of ASR by reduced phoneme set, and the design of reduced phoneme set based on LM is proposed. The experimental result shows degradations of PWCR and ASR are suppressed by the proposed approach in Section IV. Finally, conclusions and future works are described in Section V.

II. Related Works

A. Overview of Automatic Speech Recognition System

An ASR is a technique for transforming an audio signal recorded from a microphone to a text representation such as a word or a sentence referring an AM and a LM. First, the input audio signal is transformed to a sequence of feature vectors $X = x_1, x_2, ..., x_T$. A word or a sentence is decoded from

these vectors. If the target is an individual word, the ASR is called an isolated ASR, and if the target is a sentence, it is called a continuous ASR.

The continuous ASR determines the most probable sequence of words $\widehat{W} = \hat{w}_1, \hat{w}_2, ..., \hat{w}_m$ from a sequence of unknown input speech feature vectors with reference to the probability $P(W|X)$. The probability can be transformed using Bayes' rule as follows:

$$\widehat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} p(X|W)P(W) \quad (1)$$

where $W = w_1, w_2, ..., w_m$ is a sequence of words, $P(X|W)$ is likelihood given by an AM, and $P(W)$ is the probability determined by a LM.

The AM contains statistical representation of distinct sound such as phonemes. The most popular instance of the AM is hidden Markov model (HMM) whose states are given as Gaussian Mixture Model (GMM); GMM-HMM. Since 2010s, HMM with deep neural networks (DNN); DNN-HMM has seized the initiative because of a high accuracy of speech recognition [12], [13], [14], [15], [16].

The LM gives a probability to a sequence of the words $W = w_1, w_2, ..., w_m$. The most popular instance of the LM is an n-gram model which is a sequence of n words with an occurrence probability [17]. In an ASR of English, if two candidates 'fish eats plankton' and 'dish eats plankton' get highest likelihoods $P(X|W)$ by an AM, LM selects 'fish eats plankton,' because the occurrence probability $P(W)$ of the n-gram corresponding to 'fish eats plankton' is higher than the one corresponding to 'dish eats plankton.' That is to say, the most feasible sentence is recognized owing to a LM.

B. ASR for BMI

Recently, the framework of ASR has also been applied to invasive BMI. This is a next-generation BMI, which decodes the textual representation from the brain activity related to speech to assist aphasic people for the more intuitive communication.

Speech decoding from the brain activity has been studied so far [3], [4], [5], [6], [7]. In particular, Herff et al. applied an English ASR decoder to transformation of brain activity while speaking into the corresponding textual representation [3]. They trained an ECoG phoneme model as an AM in ASR. Due to the very limited amount of data to train the ECoG phoneme model, they reduced the phoneme set of size from 40 to 23 by merging phonemes together. The reduction was based on acoustic similarity of phonemes.

C. ASR for Minor Languages

The other examples of phoneme merging is the building of ASR systems for minor languages. This faces on the problem where a mass of training data cannot be collected. Some approaches merge phoneme sets of major languages to construct target minor language phoneme set to increase training data [18], [19], [20]. These approaches refer an acoustic similarity of phonemes to merge phonemes.

D. Problem of Phoneme Set Reduction

The reduction of phoneme set will increase homophones (a word that sounds the same as another but is different in spelling) in a word dictionary and will degrade the accuracy of ASR. Taking an English case for example, if two phonemes /d/ and /f/ are merged into a new phoneme, the words 'dish' and 'fish' cannot be determined from the pronunciation. According to this aspect of the phoneme reduction, merging phonemes should be performed carefully not to increase confusing homophones whose occurrence probabilities of the corresponding words in a LM are similar to each other. To our knowledge, conventional approach for the phoneme set reduction considers not the increase of homophones but only the acoustic similarity.

III. Phoneme Set Reduction Based on Language Model

In this section, a proposed design of reduced phoneme set based on a LM in continuous ASR is described. By using a LM, we define the pronunciation/word sequence confusion rate (PWCR), which estimates a degradation of an accuracy of ASR by the reduced phoneme set. In the following, the PWCR with the reduced phoneme set is first explained. Then, a reduction algorithm based on an acoustic similarity as conventional methods (Scenario 1), and the algorithm based on frequency count of phonemes in a LM (Scenario 2) are described. Finally, a reduction algorithm based on PWCR (Scenario 3) is described.

A. Pronunciation/Word Sequence Confusion Rate

As mentioned in Section II, a continuous ASR with a LM is more robust against phoneme set reduction by using word context than isolated ASR. However, it is difficult to achieve effective recognition with a smaller size of phoneme set even with use of word context.

To measure the degradation by phoneme set reduction, a PWCR is introduced in this paper. Let S be a basic phoneme set. The goal is to find of exclusive subset denoted by S_n in S such that

$$S = \bigcup_{n=1}^{N} S_n, \quad (2)$$

$$S_i \cap S_j = \phi, \quad (3)$$

where N is the number of the subsets. The subset S_n can be regarded as new phonemes in reduced phoneme set. The ASR with reduced phoneme set uses word/pronunciation dictionary described with new phonemes $S_n, n = 1, 2, ..., N$.

Using the phoneme sequence $A = S_{i_0}, S_{i_1}, ..., S_{i_{M-1}}$, where $i_m \in N, m = 0, 1, ..., M-1$, of reduced phoneme set $S = \{S_1, S_2, ..., S_N\}$, the PWCR is defined as follows:

$$\epsilon = \sum_{A} \sum_{k} (1 - p(w_k|A))p(w_k) \times 100, \quad (4)$$

where w_k is the k-th n-gram in a LM, $p(w_k|A)$ is the probability where n-gram w_k is recognized correctly from the

phoneme sequence A estimated from an AM, and $p(w_k)$ is the occurrence probability where n-gram w_k contains in LM. $p(w_k|A)$ is calculated by using Bayes' rule as follows:

$$p(w_k|A) = \frac{p(A|w_k)p(w_k)}{\sum_k p(A|w_k)p(w_k)}, \quad (5)$$

where $p(A|w_k)$ is the probability of n-gram w_k given phoneme sequence A. For example, if the phoneme sequence A corresponds to three n-grams, each probability $p(A|w_k)$ equals to $1/3$.

B. Scenario 1: Reduction Algorithm Based on Acoustic Similarity

Mak and Barnard proposed a reduction algorithm based on acoustic similarity, where Bhattacharyya distance was used for merging phonemes [21]. The Bhattacharyya distance is the similarity of two phoneme distributions in feature space. In the study [21], the Bhattacharyya distance is calculated by representing each phoneme as a Gaussian distribution. This method does not consider a LM.

C. Scenario 2: Reduction Algorithm Based on Phoneme Frequency

The first algorithm we propose is using a phoneme frequency distribution obtained from a LM. The underlying idea behind this method is that the merging among phonemes whose frequencies are less than others makes a less impact to ASR accuracy. The frequency of phoneme p_i is defined through occurrence probability of n-gram w_k in a LM as:

$$F(p_i) = \sum_k N_{p_i}(w_k)p(w_k), \quad (6)$$

where $N_{p_i}(w_k)$ is the number of phoneme p_i in the phoneme sequence A of n-gram w_k.

D. Scenario 3: Reduction Algorithm Based on PWCR

The reduction algorithms based on the acoustic similarity or phoneme frequency do not take account of the increasing of a value of PWCR. Here, we propose the second algorithm that finds reduced phoneme sets with an approximately minimum value of PWCR.

The number of ways to partition a full phoneme set of size n into k non-empty groups is known as a Stirling number of the second kind [22] denoted as:

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j \binom{k}{j}(k-j)^n. \quad (7)$$

It increases in an exponential manner as n increases. For example, if the n is 40 and the k is 20, the Stirling number $S(40,20)$ becomes approximately 10^{30}. Therefore, a brute-force search is not a realistic solution.

Due to above reason, a greedy algorithm is employed to find an approximate solution. In the first step, a problem to find a reduced phoneme set of size $n-1$ with smallest PWCR from a full phoneme set of size n is solved. This problem is solved by calculating PWCRs of all patterns of phoneme sets of size $n-1$. The number of PWCRs to be calculated is at most $\binom{n}{2} = n(n-1)/2$ that can be computed in an polynomial time. In the next step, a phoneme set of size $n-2$ with the smallest PWCR from a phoneme set of size $n-1$ found by the previous step. The number of PWCRs to be calculated in this step is $\binom{n-1}{2} = (n-1)(n-2)/2$ that is less than the previous step. Repeating this step until the phoneme set is reduced to the target size, the reduced phoneme set with approximately minimum PWCR can be found. The pseudocode of the algorithm is given in Algorithm 1.

Algorithm 1 Find reduced phoneme sets with approximately minimum PWCRs

```
Read a file "BasicPhonemeSet.txt" to BufPhonemeSet
n ⇐ GetLength(BufPhonemeSet)
for k = n − 1 to 2 do
  NewPhonemeSet ⇐ FindMinPWCRset(BufPhonemeSet)
  Write NewPhonemeSet to a file "PhonemeSet_k.txt"
  BufPhonemeSet ⇐ NewPhonemeSet
end for
```

IV. Experiment

At first in this section, the experiment setup is described. After that the reduced phoneme set obtained from the proposed algorithm (Scenario 3) is shown to have smaller PWCR than the other algorithms (Scenario 1, 2). Moreover, the reduced phoneme set is shown to prevent the ASR degradation to apply Japanese large vocabulary continuous speech recognition (LVCSR) decoder.

A. Experimental Setup

In the experiment, a corpus of spontaneous Japanese, CSJ (a large-scale database of spontaneous Japanese) [23] was used for training and evaluation of Japanese LVCSR. For training and evaluation, Kaldi [24], a free open-source toolkit for speech recognition research was used together with Kaldi-CSJ recipe[1] [25]. Kaldi-CSJ recipe has training data set for HMM with time-delay neural networks; TDNN-HMM [14], [15] as an AM and an n-gram model as a LM, and evaluation data set. Both data sets are lecture speech.

In accordance with the recipe, a TDNN-HMM was trained with the 240-hour training data set. The size of the basic phoneme set was 39 as shown in Table I. A LM was trained using a part of 450k sentences in transcription data associated with the 240-hour training data set. The 440k sentences were used for a LM training and the remaining 10k sentences were used for calculate a perplexity to evaluated the LM. The trained LM was 3-gram and Kneser-Ney discounting was applied as training options[2]. The vocabulary size was 72k. The perplexity was 69.3. In this experiment, the LM was used to calculate PWCR based on eq. (4).

[1]It is in published Kaldi code at https://github.com/kaldi-asr/kaldi/blob/master/egs/csj/s5/run.sh.

[2]The detail of the option can be referred at https://github.com/kaldi-asr/kaldi/blob/master/egs/csj/s5/local/csj_train_lms.sh.

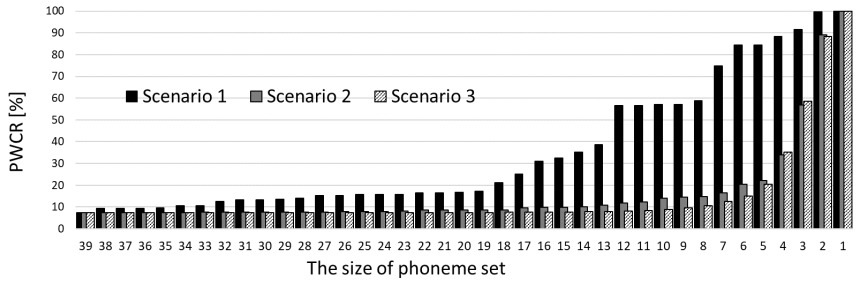


Fig. 1: A relationship between a size of reduced phoneme set and a PWCR.

TABLE I: A basic phoneme set defined in Kaldi-CSJ recipe.

Vowels (10)	a, e, i, o, u a:, e:, i:, o:, u:
Consonants (29)	b, ch, d, f, g, h, j, k, m, n, N, p, q, r, s, sh, t, ts, w, y, z, by, gy, hy, ky, my, ny, py, ry

TABLE II: WERs [%] with reduced phoneme sets

Size	Scenario	Eval1	Eval2	Eval3	AVG
39	Baseline	10.3	8.4	9.8	9.5
18	Scenario 1	15.3	12.4	15.0	14.2
	Scenario 2	11.4	9.0	10.4	10.3
	Scenario 3	10.7	8.7	10.0	9.8
10	Scenario 1	27.4	25.5	30.5	27.8
	Scenario 2	14.0	11.6	13.9	13.2
	Scenario 3	12.1	9.8	11.5	11.1

According to the recipe, standard evaluation sets labeled as Eval1, Eval2, and Eval3 in CSJ were used for evaluation. The LVCSR decoding is executed with weighted finite state transducer (WFST) [26].

B. The Comparison of Phoneme Sets by PWCR

The comparison of three reduced phoneme sets three scenarios (1, 2, and 3) by the value of PWCR. Figure 1 shows the relationship between the size of reduced phoneme set and the value of PWCR. The figure shows that as the size of the phoneme set is reduced to 18 by Scenario 1, PWCR increases over 20%. On the other hand, in Scenarios 2 and 3, when the size of phoneme set reduced to 6 and 5, PWCRs exceed 20%. This result implies that the use of a LM is effective for preventing the increase of the value of PWCR. Moreover, focusing on the reduced phoneme sets with PWCR of under 10%, the minimum size of reduced phoneme set was 14 in Scenario 2, while, 8 in Scenario 3. In this way, the proposed scenario (Scenario 3) is effective in preventing the increase of PWCR.

Figure 2 shows the behavior of the phoneme set reductions based on Scenarios 1, 2, and 3. The horizontal axis indicates phoneme symbols and the vertical axis shows the PWCR. It can be observed in Fig. 2 that more phonemes are merged with small value of PWCR in Scenario 3.

C. Speech Recognition Experiment

The reduced phoneme sets obtained from Scenarios 1, 2, and 3 were applied to Japanese LVCSR. The operation for applying reduced phoneme sets is just replacing phoneme symbols in word/pronunciation dictionary used in Kaldi-CSJ recipe and training TDNN-HMM based on the dictionary. 6 reduced phoneme sets of size 18 and size 10 from Scenarios 1, 2, and 3 were chosen to train AMs and evaluated the word error rates (WERs). Table II lists WERs of each scenario for three standard evaluation sets of CSJ: Eval1, Eval2, and Eval3. In the table. AVG denotes the average WER of these evaluation sets. It can be seen in Table II that the magnitude relationship among PWCRs of each reduced phoneme sets are maintained in terms of WERs. In comparison of reduced phoneme set of size 18 and basic phoneme set of full size 39, the degradation of WER with Scenario 3 was less than 1% for all evaluation sets. Moreover, in case of reduced phoneme set of size 10 with Scenario 3, the degradation of WER was less than 2% for all evaluation sets. These results suggest that the proposed method is very effective for Japanese LVCSR.

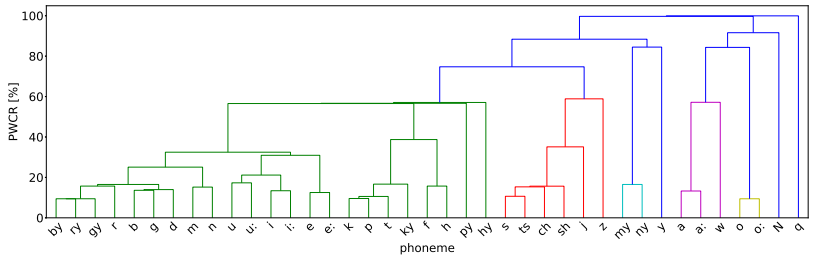
V. Conclusion

A LM-based design of reduced phoneme set for an AM was proposed. In the proposed approach, it is possible to reduce phonemes preventing degradation, regarding PWCR calculated from n-grams in a LM. In the experiment, the phoneme set designed with proposed approach was applied to Japanese large vocabulary speech recognition system. The word error rate with full 39 phonemes set was 9.5% while the error rate with the 10 phonemes set designed with the proposed approach was 11.1%. The degradation was able to be prevented within 2%.

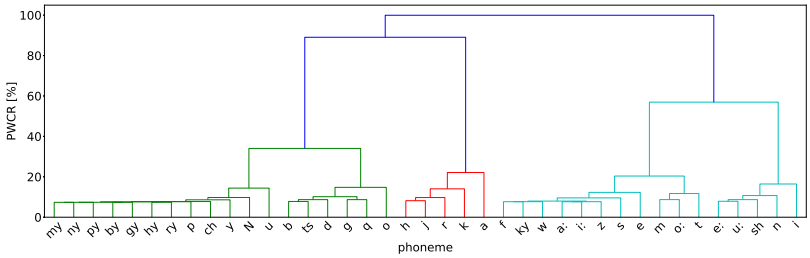
In the future works, effectiveness of the the proposed approach will be evaluated with small training data. Furthermore, the proposed approach will be applied to speech BMI.

Acknowledgments

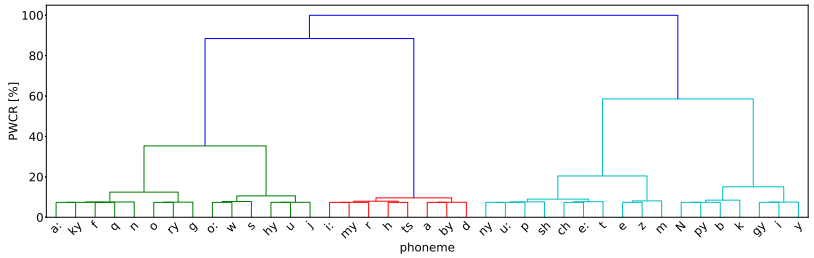
This work was supported by JST CREST (JPMJCR 1784). We thank Professor Koichi Shinoda in Tokyo Institute of Technology for giving us valuable comments.



(a) Phoneme set reduction based on Scenario 1



(b) Phoneme set reduction based on Scenario 2



(c) Phoneme set reduction based on Scenario 3

Fig. 2: A comparison among phoneme set reductions based on Scenario 1 to 3.

REFERENCES

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, 1997, pp. 1303–1306.

[3] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 9, p. 217, 2015.

[4] S. Martin, P. Brunner, I. Iturrate, J. d. R. Millán, G. Schalk, R. T. Knight, and B. N. Pasley, "Word pair classification during imagined speech using direct brain recordings," *Scientific Reports*, vol. 6, p. 25803, 2016.

[5] D. A. Moses, N. Mesgarani, M. K. Leonard, and E. F. Chang, "Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity," *Journal of Neural Engineering*, vol. 13, no. 5, p. 056004, 2016.

[6] D. A. Moses, M. K. Leonard, and E. F. Chang, "Real-time classification of auditory sentences using evoked cortical activity in humans," *Journal of Neural Engineering*, vol. 15, no. 3, p. 036005, 2018.

[7] J. A. Livezey, K. E. Bouchard, and E. F. Chang, "Deep learning as a tool for neural data analysis: speech classification and cross-frequency coupling in human sensorimotor cortex," *arXiv preprint arXiv:1803.09807*, 2018.

[8] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using htk," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 2, 1994, pp. II/125–II/128.

[9] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the Workshop on Human Language Technology*, 1994, pp. 307–312.

[10] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by mdl principle," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, 1996, pp. 717–720.

[11] T. Watanabe, K. Shinoda, K. Takagi, and E. Yamada, "Speech recognition using tree-structured probability density function," in *Third International Conference on Spoken Language Processing*, 1994, pp. 223–226.

[12] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[13] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8599–8603.

[14] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication*

Association, 2015.

[15] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[16] M. Ravanelli and M. Omologo, "Contaminated speech training methods for robust dnn-hmm distant speech recognition," *arXiv preprint arXiv:1710.03538*, 2017.

[17] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[18] M. Davel, E. Barnard, C. v. Heerden, W. Hartmann, D. Karakos, R. Schwartz, and S. Tsakalidis, "Exploring minimal pronunciation modeling for low resource languages," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] S. Hara and H. Nishizaki, "Acoustic modeling with a shared phoneme set for multilingual speech recognition without code-switching," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1617–1620.

[20] S. Sivasankaran, B. M. L. Srivastava, S. Sitaram, K. Bali, and M. Choudhury, "Phone merging for code-switched speech recognition," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, 2018, pp. 11–19.

[21] B. Mak and E. Barnard, "Phone clustering using the bhattacharyya distance," in *Fourth International Conference on Spoken Language Processing*, vol. 4, 1996, pp. 2005–2008.

[22] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics (1989)*. AIP, 1989.

[23] S. Furui, K. Maekawa, and H. Isahara, "A japanese national project on spontaneous speech corpus and processing technology," in *ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000, pp. 244–248.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, no. EPFL-CONF-192584, 2011.

[25] T. Moriya, T. Tanaka, T. Shinozaki, S. Watanabe, and K. Duh, "Automation of system building for state-of-the-art large vocabulary speech recognition using evolution strategy," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 610–616.

[26] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.