2017 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique(O-COCOSDA)
1-3 November 2017, Seoul, Korea

DEVELOPMENT OF A VIETNAMESE SPEECH RECOGNITION SYSTEM FOR VIETTEL CALL CENTER

Quoc Bao Nguyen, Van Hai Do, Ba Quyen Dam, Minh Hung Le
Cyberspace Center, Viettel Group, Vietnam
Email: {baonq2, haidv21, quyendb, hunglm11}@viettel.com.vn

ABSTRACT

In this paper, we first present our effort to collect a 85.8 hour corpus for Vietnamese telephone conversational speech from our Viettel call center. After that, various techniques such as time delay deep neural network (TDNN) with sequence training, data augmentation are applied to build the speech recognition system. Our final system achieves a low word error rate at 17.44% for this challenging corpus. To the best of our knowledge, it is the first attempt to build Vietnamese corpus and speech recognition system for the customer service domain.

Index Terms— Vietnamese speech recognition, telephone conversational speech corpus, customer service, phone call classification.

1. INTRODUCTION

Vietnamese is the sole official and the national language of Vietnam with around 76 million native speakers[1]. It is the first language of the majority of the Vietnamese population, as well as a first or second language for country's ethnic minority groups.

At the early time, there were several attempts to build Vietnamese large vocabulary continuous speech recognition (LVCSR) system where most of them developed on read speech corpuses [1-4]. In 2013, the National Institute of Standards and Technology, USA (NIST) released the Open Keyword Search Challenge (Open KWS), and Vietnamese was chosen as the "surprise language". The acoustic data are collected from various real noisy scenes and telephony conditions. Many research groups around the world have proposed different approaches to improve performance for both keyword search and speech recognition [5-7].

In this paper, we present our effort to collect a Vietnamese corpus and build a LVCSR system for our customer service call center. To the best of our knowledge, there is no such type of speech corpus for Vietnamese as well as no LVCSR system for this domain. After that a text classifier is place on the top of speech recognition for phone call classification. The output of the system is used for customer service management purposes. Our premium goal is to build a fully automatic call center in the future. This is a very important task with a huge market. For example, only in Viettel Telecom, our call center receives around 500,000 phone calls every day.

To build a speech recognition system, we collect 85.8 hours audio data from our call center. Transcription is generated by our 400 call center staffs i.e., agents. Various techniques are applied such as time delay neural network (TDNN) [8] with sequence training, data augmentation [9], etc. Finally, we achieve 17.44% word error rate for this challenging task.

The rest of this paper is organized as follows: Section 2 describes our corpus. Section 3 gives a description of the proposed system. Section 4 presents experimental setup and results. Discussion is shown in Section 5 and we conclude in Section 6.

2. CORPUS DESCRIPTION

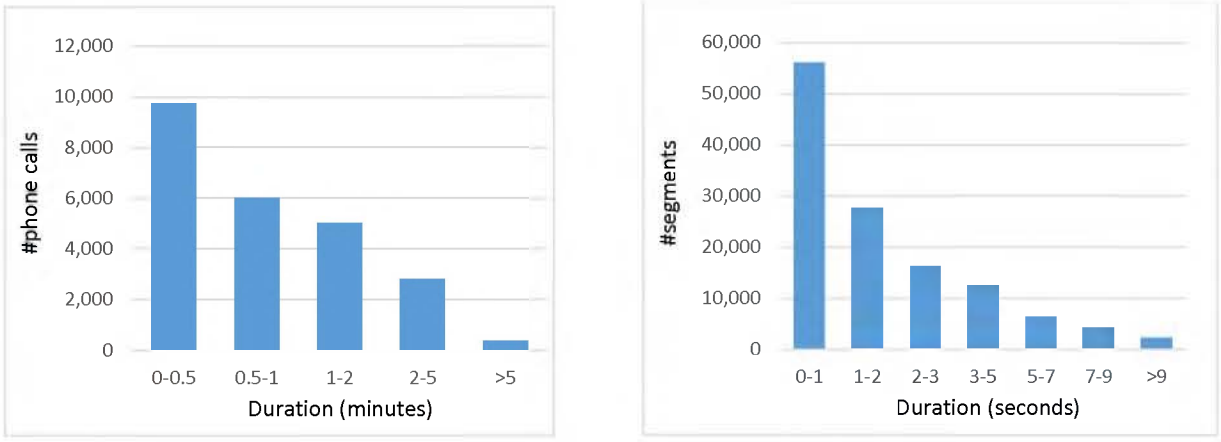
We collect phone calls from the Viettel customer service call center. The sampling rate is 8kHz, with a resolution of 8 bits/sample. In the corpus, there are 50 agents which served totally 23,932 phone calls. Two channels i.e., the agent and the customer channels are separately recorded. After that, a voice activity detection (VAD) module is applied to segment 23,932 phone calls into 153,825 speech segments with a total duration of 85.8 hours. To achieve the transcription for each segment, 400 transcribers from the Viettel customer service department are hired[2]. To improve the transcription quality, each segment is transcribed by two transcribers. After that, we highlight the differences between the two transcriptions and ask one more person to verify and correct that segment. The transcription contains not only the text transcription, but also noise and silent tags which are labelled as laughter, applause, silence.

Figure 1(a) shows the distribution of phone call durations. We can see that most of phone calls have a short duration. The average length is 1.0 minute. Figure 1(b) presents the distribution of segment lengths generated by the VAD module. The average length of each segment is 2.0 seconds.

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] We hired transcribers from our customer service department instead of professional transcribers since there are a lot of technical terms in the phone calls. Hence, "out-domain" transcribers may have difficulty to transcribe.

978-1-5386-3333-5/17/$31.00©2017 IEEE



(a) Phone call duration (b) Segment duration

Figure 1. The distribution of phone call and segment durations.

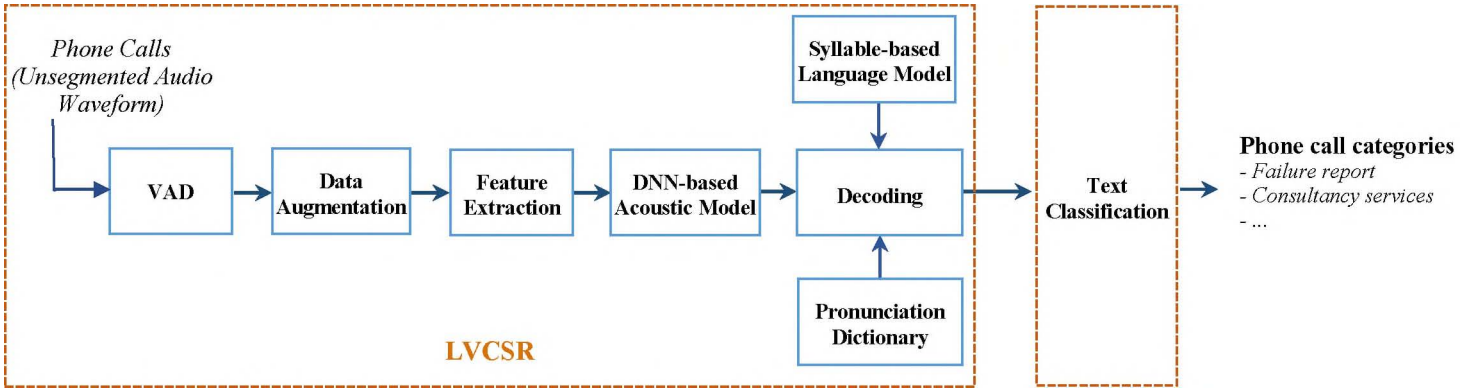


Figure 2. The proposed system for phone call classification.

3. SYSTEM DESCRIPTION

Figure 2 illustrates the proposed system. We first build a LVCSR system and then place a text classifier on the top for phone call classification. Specifically, audio waveform from phone calls is first segmented with a voice activity detector (VAD). To increase the data quantity, data augmentation is adopted. Feature extraction is then applied to use for the acoustic model. For decoding, acoustic model is used together with syllable-based language model and pronunciation dictionary. After decoding, recognition output is used to classify phone calls into different groups. In the next subsections, the detailed description of each module is presented.

3.1. Voice activity detection

In our call center, the agent channel and the customer channel are separately recorded. Hence, there are a lot of silent in each audio channel and they need to be divided into short sentence-like segments. In order to detect voice activity and segment the audio, we use 10 hours of data to train a VAD model. After that this model is used to align the audio. The segmented audio is then generated by cutting adjacent non-speech phoneme (noise and silent) which more than duration threshold (1 second) in aligned output.

3.2. Data augmentation

To build a reasonable acoustic model, hundreds to thousands hours of audio are needed. However, to achieve transcribed audio data is very costly. To overcome this, many techniques have been proposed such as semi-supervised training [10], phone mapping [11], exemplar-based model [12], mismatched crowdsourcing [13]. In this paper, we use a simple approach called data augmentation. It is a common strategy adopted to increase the data quantity to avoid overfitting and improve the robustness of the model against different test conditions [14]. In this study, we increase training data size using a data augmentation technique called audio speed perturbation [9]. Speed perturbation produces a warped time signal, for example, given speech waveform signal $x(t)$, time warping by a factor

α will generate signal $x(\alpha t)$. In this study, we use three different values of α i.e., 0.9, 1.0, 1.1.

3.3. Feature extraction

We use Mel-frequency cepstral coefficients (MFCCs) [15], without cepstral truncation are used as input feature i.e., 40 MFCCs are computed at each time step which is similar setup in [8]. Since Vietnamese is a tonal language, pitch feature is used to augment MFCC. The effect of pitch feature is examined in the experiment section.

3.4. Acoustic model

Two advanced acoustic models are considered in this paper i.e., Gaussian mixture model with speaker adaptive training [16] (GMM-SAT) and time delay deep neural network (TDNN) with sequence training [8].

3.5. Pronunciation dictionary

Vietnamese is a monosyllabic tonal language. Each Vietnamese syllable can be considered as a combination of initial, final and tone components. Therefore, the pronunciation dictionary (lexicon) needs to be modelled with tones. As in [17], we use 47 basic phonemes. Tonal marks are integrated into the last phoneme of syllable to build the pronunciation dictionary for 6k popular Vietnamese syllables.

In order to build the dictionary for foreign and technical words, we select 5k popular foreign words from web newspapers together with 500 words in the customer service domain. These words are then manually pronounced in the Vietnamese pronunciation. For generating unknown words in training data, we employ grapheme-to-phoneme (G2P) conversion using Sequitur G2P open-source toolkit [18] trained on the 5k foreign lexicon. As a result, the total number of words in our lexicon is about 12k words. This lexicon is used for training as well as decoding.

3.6. Language model

A syllable-based language model is built from training transcription. 4-gram language model with Kneser-Ney smoothing is used after exploring different configuration. We also tried to enlarge the text corpus by using different text sources such as from web text or movie closed caption, however no improvement is observed. A possible reason is that those text sources are too different from the customer service domain.

3.7. Text classification

After decoding, recognition output is used for text classification to classify phone calls into different groups such as failure report, consultancy services. In this preliminary study, we simply classify the phone calls based on a keyword list. Specifically, each group has a list of keywords defined by our customer service department. After decoding, a keyword detector finds the keywords from the decoding output (1-best). Each keyword is simply assigned an equal score. A phone call will be classified to the group which has the highest score.

4. EXPERIMENTS

4.1. Experimental setup

We first define the training and the test sets from the corpus. We extract 19,672 phone calls from 43 agents to form the training set. The training set length is 70 hours with 125,337 segments. The remaining set consists of 4,260 phone calls from 7 agents is used for the test set. The test set duration is 15.8 hours with 28,488 segments. With this setup, there is no overlapped speaker between training and the test sets.

The Kaldi speech recognition toolkit [19] is used to build speech recognition. SRILM toolkit [20] is used to build language model. Performance of all the systems is evaluated in word error rate (WER).

4.2. Acoustic feature

We first evaluate the system with different types of input feature. MFCC features are used in our baseline system for the acoustic model. In addition, we evaluate system performance when MFCC is augmented with pitch feature. As shown in the middle column of Table 1 using pitch feature results in a significant word error rate (WER) reduction (from 37.38% to 31.15%)

Table 1. Word error rate (%) of the speech recognition system using two different input features with two different types of pronunciation dictionary.

Feature	Non-tonal dictionary	Tonal dictionary
MFCC	37.38	36.72
MFCC+pitch	31.15	28.99

4.3. Pronunciation dictionary

In Section 4.2, the non-tonal dictionary is used i.e., all entries in the dictionary are pronounced as a sequence of non-tonal phonemes (phonemes without tonal marks). Vietnamese is a tonal language hence obviously using tonal phone set is an appropriate choice. As comparing the middle and the last columns of Table 1, we see that using the tonal dictionary can significantly improve the recognition performance. Also note that for the case when MFCC is used as the input feature, we obtain only 0.65% improvement by using the tonal dictionary. In contrast, when MFCC is augmented with pitch, the improvement by

using the tonal dictionary is significantly larger (2.16%). It shows that using tonal dictionary especially improves recognition performance when input feature is augmented with tonal feature i.e., pitch.

4.4. GMM vs. DNN acoustic models

In the previous experiments, the GMM acoustic model with speaker adaptive training (SAT) was used. This section investigates the DNN-based acoustic model. The advantages of DNN over GMM for acoustic modelling have been shown by many researchers [21, 22]. In this paper, we use a variation of DNN called time delay neural network (TDNN) proposed recently [8]. The middle column of Table 2 shows WER% of our system with different types of acoustic model. Note that in this case, input feature is MFCC+pitch and the tonal dictionary is used. We first use TDNN with frame-based cross-entropy training criterion (TDNN1). It is seen that WER drops significantly (from 28.99% to 20.20%) by using TDNN for acoustic modelling. After that TDNN with sequence training (TDNN2) is applied based on a state-level variant of the Minimum Phone Error (MPE) criterion, called sMBR [23]. By using sequence training, we achieve around 2% further improvement. With more sMBR iterations, better performance is achieved. WER seems saturated after 4 iterations.

Table 2. Word error rate (%) of speech recognition system using GMM and DNN acoustic models without and with data augmentation.

Acoustic model		w/o data augmentation	with data augmentation
GMM-SAT		28.99	27.92
TDNN1 (w/o sMBR)		20.20	19.18
TDNN2 (w/ sMBR)	Iteration 1	18.34	17.41
	Iteration 2	18.19	17.44
	Iteration 3	18.06	17.31
	Iteration 4	18.04	17.28

4.5. Data augmentation

To increase the training data size, we adopt a data augmentation technique called audio speed perturbation [9]. In this study, three versions of the original speech signal, $x(t)$ are created i.e., $x(0.9t)$, $x(t)$, $x(1.1t)$. After that, feature extraction is applied on the new speech signal to train the acoustic model as in the conventional way.

The last column of Table 2 shows WER% of different acoustic models after applying data augmentation. It is clearly seen that using data augmentation consistently reduces WER from 0.75% to 1.07% for different acoustic models.

5. DISCUSSION

In this paper, various techniques have been applied to improve speech recognition performance for customer service telephone calls. The final system achieves a comparative WER of 17.44%. This is a reasonable number for this challenging telephone conversational speech corpus. For analysis, we breakdown performance of our system for customer and agent sides (Table 3). We realize that for agent side, we achieve a much better performance than the customer side. It can be explained that the speech quality our customer service staff (agent) is much better than the customers' one for example less noise. In addition, spoken language uttered by our staff is more formal and hence the language model is easier to capture it.

Table 3. Word error rate (%) breakdown for agent and customer sides.

Agent+Customer	Agent	Customer
17.44	10.29	26.14

We also realize that our recognition system does not performs well for the central and southern dialects. The reason is that the training corpus was collected by our call center located in the Hanoi (north of Vietnam).

6. CONCLUSION

In this paper, we presented the effort to develop a Vietnamese speech recognition system for our phone call classification purpose to improve customer service management. Various techniques have been applied such as TDNN with sequence training, data augmentation to significantly improve speech recognition performance. Our best system achieved a comparative 17.44% WER. In the future, we will collect data not only from northern dialect but from central and southern dialects to make our system is more robust with different dialects in Vietnam.

7. REFERENCES

[1] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, and John-Paul Hosom, "Vietnamese large vocabulary continuous speech recognition," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2005, pp. 492–495.

[2] Quan Vu, Kris Demuynck, and Dirk Van Compernolle, "Vietnamese automatic speech recognition: The flavor approach," in Proc. *the 5th International Conference on Chinese Spoken Language Processing (ISCSLP)*, 2006, pp. 464–474.

[3] Tuan Nguyen and Quan Vu, "Advances in acoustic modeling for Vietnamese LVCSR," in Proc. *International Conference on Asian Language Processing (IALP)*, 2009, pp. 280–284.

[4] Ngoc Thang Vu and Tanja Schultz, "Vietnamese large vocabulary continuous speech recognition," in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

[5] Nancy F. Chen, Sunil Sivadas, Boon Pang Lim, Hoang Gia Ngo, Haihua Xu, Bin Ma, and Haizhou Li. "Strategies for Vietnamese keyword search," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4121-4125.

[6] Tsakalidis, Stavros, Roger Hsiao, Damianos Karakos, Tim Ng, Shivesh Ranjan, Guruprasad Saikumar, Le Zhang, Long Nguyen, Richard Schwartz, and John Makhoul. "The 2013 BBN Vietnamese telephone speech keyword spotting system," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7829-7833.

[7] I-Fan Chen, Nancy F. Chen, and Chin-Hui Lee, "A keyword-boosted sMBR criterion to enhance keyword search performance in deep neural network based acoustic modeling," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH),* 2014.

[8] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 3214-3218.

[9] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH),* pp. 3586-3589, 2015.

[10] Haihua Xu, Hang Su, Eng Siong Chng, and Haizhou Li. "Semi-supervised training for bottle-neck feature based DNN-HMM hybrid systems," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH),* 2014.

[11] Van Hai Do, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Context-dependent phone mapping for LVCSR of under-resourced languages," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 500–504.

[12] Van Hai Do, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Kernel Density-based Acoustic Model with Cross-lingual Bottleneck Features for Re-source Limited LVCSR," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 6–10.

[13] Van Hai Do, Nancy F. Chen, Boon Pang Lim, and Mark Hasegawa-Johnson, "Speech recognition of under-resourced languages using mismatched transcriptions," in Proc. *International Conference on Asian Language Processing (IALP)*, 2016, pp. 112–115.

[14] Navdeep Jaitly and Geoffrey E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in Proc. *ICML, Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[15] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[16] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, pp. 1043-1046.

[17] Quoc Bao Nguyen, Tat Thang Vu, and Chi Mai Luong, "The Effect of Tone Modeling in Vietnamese LVCSR System," Procedia Computer Science 81 (2016): 174-181.

[18] Maximilian Bisani and Hermann Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," Speech Communication, Volume 50, Issue 5, May 2008, Pages 434-451

[19] Daniel Povey, et al. "The Kaldi speech recognition toolkit," in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[20] Andreas Stolcke et al., "SRILM-an extensible language modeling toolkit," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH),* 2002.

[21] Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine 29.6 (2012): 82-97.

[22] Van Hai Do, Xiong Xiao, and Eng Siong Chng, "Comparison and Combination of Multilayer Perceptrons and Deep Belief Networks in Hybrid Automatic Speech Recognition Systems," in Proc. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2011.

[23] Matthew Gibson, "Minimum Bayes risk acoustic model estimation and adaptation," Ph.D. dissertation, University of Sheffield, 2008.