VAIS-Speech: An Overview of Automatic Speech Recognition and Text-to-speech Development at VAIS

Quoc Truong Do
Vietnam Artificial Intelligent System
Email: support@vais.vn

***Abstract*—In this paper, we describe the development of automatic speech recognition (ASR) and text-to-speech (TTS) systems at VAIS. Our speech engines utilized the state-of-the-art technologies that have been used in many popular languages such as English and Japanese. Moreover, we also designed many features of the core engine that are highly optimized for Vietnamese. We evaluated our ASR engine on large test sets including news (9.3 hours) and mobile phone (10 hours) domains spoken by both Northern and Southern accents. As the result, we achieved 5.58% WER on the news test set and 7.99% on the mobile phone set.**

I. Introduction

Automatic speech recognition (ASR) and speech synthesis technologies are two important components for a speech based human-machine interaction. Both ASR and TTS are active research areas that has been well studied for few decades. And thank to the powerful of deep learning, recently, these technologies are deployed in many softwares and companies making the human-machine communication at remarkable high accuraries. In this paper, we present our deep learning based ASR and TTS engines for Vietnamese language. Our ASR engines can recognize words spoken by both north, south and center regions of Vietnam and the TTS engines support both north and south accents.

II. Automatic speech recognition

In a conventional Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) acoustic model, the state emission log-likelihood of the observation feature vector o_t for certain tied state s_j of HMMs at time t is computed as

$$\log p(\mathbf{o}_t|s_j) = \log \sum_{m=1}^{M} \pi_{jm} \mathbb{N}(\mathbf{o}_t|s_j) \tag{1}$$

where M is the number of Gaussian mixtures in the GMM for state j and π_{jm} is the mixing weight. As the outputs from DNNs represent the state posteriors $p(s_j|\mathbf{o}_t)$, a DNN-HMM hybrid system uses pseudo log-likelihood as the state emissions that is computed as

$$\log p(\mathbf{o}_t|s_j) = \log p(s_j|\mathbf{o}_t) - \log p(s_j), \tag{2}$$

where the state priors $\log p(s_j)$ can be estimated using the state alignments on the training speech data. In our experiment, we used an deep neural network (DNN) with 7 layers to generate the state posteriors $p(s_j|\mathbf{o}_t)$. Our language model use 4-gram model and it is trained from news and conversation dataset.

III. Text-to-speech

A. Speech concatenation approach

Speech concatenation [1] is an approach that synthesize audios by concatenate speech segments from a database. The idea is as follows, first, audio and text are aligned at the phoneme level to provide information about where the phoneme is located in the speech sound. Second, the text is also analyze to provide linguistic information, such as phonetic context, part-of-speech tags, word position. Finally, a database is built which contains all speech segments and phonetic information.

At the synthesis time, the speech segment is chosen to minimize the combination of the target cost which measured by a heuristic distance between contexts and the concatenation cost which measured by speech parameter distortion (Fig. 1).

Because the audio is synthesized using the speech segment extracted directly from original audios, it provides very high quality speech waveform signal. If the model can choose correctly speech segments, it is very difficult for human to distinguish the synthetic voice and the natural voice. However, allthough the concatenation approach can produce high quality speech waveform, it comes with some disadvantages. First, it has large footprints because all speech segments are stored in the model. Our actual model size is approximately 1GB when trained on 6k utterances. Second, the sound sometime has unstable quality due to wrong alignment and mistake during segment selection.

B. HMMs approach

The HMM approach models speech by using Gaussian mixture models. Each phoneme is modeled by an HMM instead of many speech segments in concatenation approach. At the synthesis time, a sentence HMM is constructed from the given text. Then the speech parameter is predicted to maximize the likelihood probability,

$$\hat{\boldsymbol{O}} = \underset{\boldsymbol{O}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, T), \tag{3}$$

where $\boldsymbol{O}$ is speech parameters, λ is the model parameters, and T is the length of speech that we want to generated.

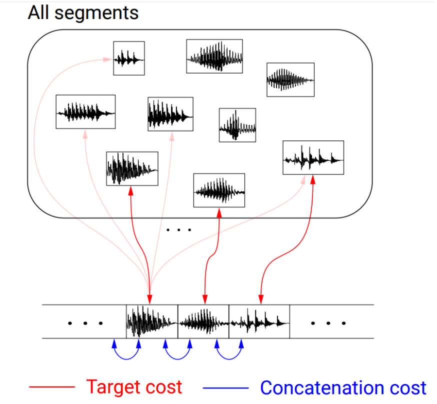

Figure 1. Speech concatenation procedure.

Unlike unit concatenation approach where we need to collect large amount of speech data to have good quality, the HMM approach can trained a model with just few hundreds phonetic-balanced utterances. This allows us to quickly train a fairly good model given small amount of data.

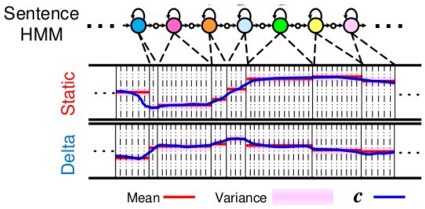

Figure 2. HMM-based speech synthesis

The HMM voice has very low footprints, in fact, our model trained on 6k utterances takes only 5MB of storage and 10MB of memory when fully loaded. The approach can also provide very flexible voices by changing model parameters [2], and also be able to adapted to someone else voices with minimal data collection required[3].

C. Front-end text processing

Vietnamese is a complex language where one word can be pronounced in different ways depending on the context. Another problem is abbreviations that is very often being used in newspaper, such as TPHCM, VKSND. The list of abbreviation is endless and have no rules to pronounced.

To make the speech synthesis more useful for general tasks, we define a set of regression rules for date, numbers, date-of-birth, time, units (such as currencies, temperature, weights), and develop a toolkit for define abbreviation words. The processing procedure is illustrated in Fig. 3.

D. Available voices

At the current stage, we provide 2 speech concatenation female voices with Northern and Southern accent and 1 HSMM voice. All voices are made available for online access and the HSMM voice is also available for offline access including PC and mobile platforms.

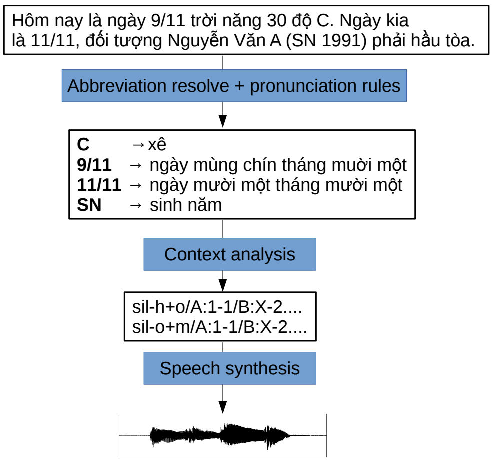
Hôm nay là ngày 9/11 trời nắng 30 độ C. Ngày kia là 11/11, đối tượng Nguyễn Văn A (SN 1991) phải hầu tòa.
Abbreviation resolve + pronunciation rules
C → xê
9/11 → ngày mùng chín tháng mười một
11/11 → ngày mười một tháng mười một
SN → sinh năm
Context analysis
sil-h+o/A:1-1/B:X-2....
sil-o+m/A:1-1/B:X-2....
Speech synthesis

Figure 3. The text analysis front-end procedure.

Table I
Word Error Rate (WER) Evaluation on news and mobile test set

Model	News	Mobile
DNN 7x1024	5.58	7.99

The dataset used for training models is described as follows:

- **Southern accent:** 5.8k utterances of female voice collected from VOV audio speech. The length of each utterance varies from 5 to 15 words.
- **Northern accent:** 6k utterances of female voice. The length of each utterance varies from 14 to 17 words.

IV. Experimental and Results

A. ASR evaluation

In this section, we describe our ASR training setup and evaluation. Our acoustic model is trained using 1200 hours including North, South and Central regions. Audio speech signals are sampled at 16kHz sampling frequency. The acoustic feature includes both Mel Frequency Ceptral Coefficients (MFCC) feature and pitch feature. The MFCC feature is extracted using an windown of 25ms and a shift of 10ms.

In our experiments, a deep neural network (DNN) with 7 layers is used to generate the state posteriors. Each layer has 1024 nodes. To train the DNN, the cross-entropy objective funtion is used. We used the 4-gram model language model in which has vocabulary size of 6500 words. And the output lattices is re-scored with a 5-gram language model.1

The result is shown in Fig. I. As we can see, our sytem perform very well on the news dataset when the audio quality is good and clean. While on the mobile test set where the audio is noisy and include more conversation and fast-speaking style, the performance is dropped by approximately 2%.

B. TTS training setup

Our TTS system is trained on 5,000 utterances spoken by a female speaker with Northern accent using the HTS toolkit. Speech features are 40 dimension mel-generalized cepstral coefficients (MGC), 1 dimension band apperiodic feature and 1 dimension log F0 extracted with a windows of 25 frames and 5 frames shift.

V. Conclusion

In this paper, we briefly describe our ASR and TTS development. Both engines achieved high performance and can be used in general domain. The ASR engine works well not only under clean but also very noisy and far-field environment. More detail and desmonstration are available on our website: https://vais.vn.

References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *In Proceeding of ICASSP*, 1996, pp. 373–376.

[2] N. Takashi, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE*, vol. 90, no. 9, pp. 1406–1413, 2007.

[3] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE*, vol. 90, no. 2, pp. 533–543, 2007.