Developing forced alignment module for large corpus of Broadcast news

Red Hen Labs

Abstract

In this paper we explain the steps for building forced alignment module given a broadcast speech corpus. We use acoustic and various language models to obtain phone and word labels and timestamps for the videos. To prune noisy/clipped/disfluent/unaudible/unintelligible/mislabeled speech, confidence measure based on the estimate of posterior probability is used.

Index Terms: Unit selection synthesis, automatic speech recognition, audiobooks, confidence measures.

1. Introduction

In this paper we use free audiobooks available at Librivox (librivox.org) for developing the alignment system. These audiobooks contain read speech which are recorded books from project Gutenberg (gutenberg.org). We use these audiobooks without corresponding text, and try to obtain accurate text (transcripts and timestamps) using open source Librispeech [4] acoustic and language model (LM) (available at kaldi-asr.org). Generally, the goal of a large vocabulary continuous speech recognizer is to hypothesize the actual spoken/intended word sequence even when the speech is corrupted by noise to such an extent that it is unintelligible to a human listener. Use of noise suppression techniques and noise robust acoustic model trained on large amount of noisy data, and more importantly use of a large language model prepared on a billion words help to acheive the goal [6]. Section 3.1 walks the steps to obtain such a word sequence. We don't use a simple unigram LM having equal unigram probability for each in-vocabulary word for making the recognition independent of language model probabilities, and disadvantages of this approach [7] are discussed in section 3.1.

LM plays an important role in the search space minimization and also on the quality of the competing alternative hypotheses in the lattice. Here, Librispeech LM is used as the task is to decode audiobooks available at Librivox. The Librispeech LM won't be that useful if we're to decode audio of a lecture or a speech which encapsulates a specific topic and vocabulary. In such a case, it would be better to prepare a new LM based on the text related to the topic in the audio and interpolate it with LM prepared on Librispeech corpus.

Posterior probability, by definition, tells the correctness or confidence of a classification. In speech recognition, posterior score reflects the goodness of fit with the language model too, in addition to that with acoustic model. Hence, we consider the posterior score after the influence of LM is nullified, and when the posterior score conveys only the acoustic match. It is also true that posterior score depends on the acoustic model, (mis)match between the training and test conditions, and also on the error rate [8]. To alleviate the dependence of the posterior on one signal acoustic model, we re-compute the posteriors after rescoring the same lattices using a different acoustic model such as one trained on articulatory features. Such system combination makes the posteriors more reliable.

2. Using Librispeech models for decoding

Librispeech is a fairly recently made available largest (~1000 hours) open source continuous speech corpus in English. It is a collection of parts of several audiobooks downloaded from Librivox. It mainly consists of two parts: 460 hours of clean data, and 500 hours of data containing artificially added noise. Since our goal is to decode audiobooks and consequently develop a USS system, using models trained on Librispeech data seemed to be the best choice.

Each broadcast news item consists of an approximate transcription file and a video file from which the audio needs to be extracted. The audio files were downloaded and converted to 16kHz WAV format. These wavefiles were then chopped based on silence intervals of 0.3 seconds or more to create phrasal chunks averaging 15 seconds in length. The chunks were power-normalized. An average chunk length of 15 seconds is sufficient enough to capture intonation variation, and does not create memory shortage problems during Viterbi decoding. It is also observed that decoding is faster and more accurate compared to when it is performed for much longer chunks.

3. Experiments and Evaluation

In this section, we first explain the steps to obtain accurate phone hypothesis and timestamps for audiobooks, and then to develop a unit selection voice.

3.1. Obtaining accurate hypothesis and timestamps

Our goal is to obtain accurate labels and timestamps. To prune out noisy/disfluent/unintelligible regions from audio, we also need a confidence (or posterior probability) score that reflects the acoustics reliably. In addition the confidence score should not reflect language model score, and instead should reflect purely acoustic likelihood score.

1. **Feature extraction**: The first step is to extract features from audiobooks. 39 dimensional acoustic feature vectors (12 dimensional MFCC and normalized power, with their deltas and double-deltas) are computed. Cepstral mean and variance normalization is applied. The feature vectors are then spliced to form a context window of seven frames (three frames on either side) on which linear and discriminative transformation such as Linear Discriminant Analysis is applied which helps achieve dimensionality reduction.
2. **Decoding the audiobook using speaker adapted features**: We use the p-norm DNN-HMM speaker independent acoustic model [11] trained on 460 hours of clean data from Librispeech corpus for decoding. The decoding is carried out

in two passes. In the first pass, an inexpensive LM such as pruned trigram LM is used to constrain the search space and generate a lattice for each utterance. The alignments obtained from the lattices are used to estimate speaker dependent fMLLR or feature-space MLLR transforms. In the second pass of decoding, an expensive model such as unpruned 4-gram LM is used to rescore the lattice, and obtain better LM scores.

Combination of phone and word decoding: The lattices generated in the previous step don't simply contain word hypotheses, but instead contain a combination of phone and word hypotheses. Phone decoding, in tandem with word decoding helps reduce errors by a significant proportion in the occurrence of out-of-vocabulary words or different pronunciations of in-vocabulary words. A combination of phone and word decoding can be performed by simply including the phones in the text from which the LM is prepared. An example to highlight the use of this technique is as below. We can observe the sequence of phones hypothesized because of the difference in pronunciations of the uttered word and its pronunciation in the lexicon.

Example:

Bayers	b ey er z (pronunciation in the lexicon)
Reference:	Performed by Catherine Bayers
Hypothesis:	Performed by Catherine b ay er z

3. **Improving decoded transcripts using LM interpolation**: We find the 1-best transcripts from the lattices generated in the previous step. These 1-best transcriptions encapsulate the *specific* vocabulary, topic and style of the book. As a result, a LM computed purely from the decoded text is expected to be different and more relevant for recognition compared to a LM prepared from all Librispeech text. We exploit this fact to further improve the decoding by creating a new LM which is a linear interpolation of LM prepared on decoded text and LM prepared on entire Librispeech text. The LM interpolation weight for the decoded text is set to 0.9 to create a strong bias towards the book text. A new lattice is then generated using the new LM.

4. **Nullifying LM likelihood before computing posteriors**: [7] Our goal is to obtain a 1-best hypothesis and associated posterior scores that match well with the acoustics, and have little or no influence of language on them, for USS task. Every word/phone hypothesis in a particular lattice of an utterance carries an acoustic and language model likelihood score. In a pilot experiment, we tried generating a lattice based upon pure acoustic score in the following way. We prepared a unigram LM from text containing a unique list of in-vocabulary words and phones. Just one single occurrence of each word and phone in the text made sure that frequency, and consequently the unigram probability of each word and phone is the same. It was found that the 1-best hypothesis produced by this method was nowhere close to the reference sequence of words. This outcome was understandable as we had not put any language constraints, and the decoder was tied down to choose between several words (~200,000) in the lexicon just on the basis of acoustics. Understandably, the phone hypothesis obtained by lexicon look up was also worse. The example below demonstrates the large error in the hypothesis when a unigram LM was used.

   Reference: RECORDING BY RACHEL FIVE LITTLE PEPPERS AT SCHOOL BY MARGARET SIDNEY.

   Unigram LM: EDINGER RACHEL FADLALLAH PEPPERS SAT SQUALL PRIMER GRITZ SIDNEY.

   We therefore resorted to the following approach. Rather than using the above mentioned unigram LM from the start, i.e. for the generation of lattices, it proved to be more useful to rescore the lattices (obtained in the previous step after LM rescoring and LM interpolation) containing alternative hypotheses which are much closer to the sequence of reference words. The posteriors, therefore obtained also reflected pure acoustics. The sentences below show the 1-best output of the lattice in the previous step (after performing LM rescoring and interpolation), and the 1-best output after rescoring the same lattice with unigram LM having equal unigram probabilities for each in-vocabulary word.

   4-gram LM: READING BY RACHEL FIVE LITTLE PEPPERS AT SCHOOL BY MARGARET SIDNEY.

   Nullified LM: READING *MY* RACHEL *SIL FILE IT ILL* PEPPERS AT SCHOOL BY MARGARET SIDNEY.

   The second hypothesis is more close to acoustics. Difference between the two hypotheses are italicized. It is clear that the 1-best transcription is better and also closer to acoustics when the unigram LM is used for rescoring the lattice generated in the previous step, rather than using it for generating the lattice from scratch. Consequently, the phone-level transcripts are also better, and the posteriors purely reflect acoustic match.

5. **Articulatory rescoring and system combination**: The lattices generated in the previous step are rescored using a p-norm DNN-HMM acoustic model, trained on articulatory features, and speaker adapted articulatory features to yield a new lattice. This new rescored lattice is then combined with the original lattice to form a combined lattice. Pure articulatory feature based recognition is not as robust, and hence lattices are not generated using the acoustic model trained on articulatory features, and it is rather used for rescoring the lattice generated using acoustic model trained on MFCC. Lattice combination provides the advantage that two lattices scored with two different models and features contain complementary information, which yields a lattice with more robust acoustic scores. The 1-best hypothesis obtained from the above lattice is also more accurate. Word lattices are then converted to phone lattices. The 1-best phone sequence from the phone lattice alongwith the posteriors is what we use for building USS system.

3.1.1. Data pruning using phone confidence measure

There is a need to prune all such instances of bad data as USS is very sensitive to noisy data and transcriptions, and its success depends on having noise-free audio containing fluent speech and accurate labels and timestamps. I We use confidence measure based on the estimate of posterior probability to prune the bad data. Posterior probability, by definition, tells the correctness or confidence of a classification. In speech recognition, the posterior probability of a phone or a word hypothesis w given a sequence of acoustic feature vectors $O_1^T = O_1 O_2 \, .. \, O_T$ is computed (as given in equation below) as the likelihood of all paths passing through the particular phone/word (in around same time region) in the lattice normalized by the total likelihood of all paths in the lattice. It is computed using forward-backward algorithm over the lattice. Posterior probability computed from lattices outperforms the many confidence measures proposed in literature [14].

Let W_s and W_e respectively denote word sequences preceding and succeeding word w whose posterior probability is to be computed. Also let $W^{'}$ denote the word sequence ($W_s w W_e$). Then,

$$p(w|O_1^T) =$$

$$= \sum_{W_s}\sum_{W_e} p(W_s w W_e|O_1^T)$$

$$= \frac{\sum_{W_s}\sum_{W_e}\left[p(O_1^{t_s}|W_s)p(O_{t_s}^{t_e}|w)p(O_{t_e}^{T}|W_e)p(W^{'})\right]}{p(O_1^T)}$$

$$= \frac{\sum_{W_s}\sum_{W_e}\left[p(O_1^{t_s}|W_s)p(O_{t_s}^{t_e}|w)p(O_{t_e}^{T}|W_e)p(W^{'})\right]}{\sum_{W}\left[\sum_{W_s}\sum_{W_e}\left[p(O_1^{t_s}|W_s)p(O_{t_s}^{t_e}|w)p(O_{t_e}^{T}|W_e)p(W^{'})\right]\right]}$$

In above equation, $p(O_1^T)$ in the denominator is approximated as the sum of likelihoods of all paths in the lattice. As can be seen, the posterior score has contribution from LM too (the term $p(W^{'})$ in the numerator which signifies LM likelihood). Hence, we consider the score after contribution of LM is nullified (as explained in section 3.1), and when the posterior score purely reflects the acoustic match. It is also true that posterior score depends on the acoustic model, (mis)match between the training and test conditions, and also on the error rate [8]. To alleviate the dependence of the posterior on one signal acoustic model, we re-compute the posteriors after rescoring the same lattices using a different acoustic model such as one trained on articulatory features (explained n section 3.1). Such system combination makes the posteriors more reliable.

4. Conclusions

In this paper, we saw that a natural sounding unit selection voice can be developed given a single speaker audio, which could be lecture, read or spontaneous speech, without corresponding text. Steps to obtain accurate phone labels (which reflect acoustics) and associated timestamps are described. Confidence measure based on the estimate of posterior probability is used to prune noisy/disfluent/unaudible/unintelligible/clipped speech. Suitability of using contextual units such as quinphone given large amount of prosodically rich data is discussed. Effectiveness of WSOLA algorithm to smoothen the signal near joins and increase naturalness is demonstrated.

5. References

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

[2] M. Bulut, S. S. Narayanan, and A. K. Syrdal, “Expressive speech synthesis using a concatenative synthesizer.,” in *Proc. of Interspeech*, 2002.

[3] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, “A corpus-based approach to expressive speech synthesis,” in *Proc. of Fifth ISCA Workshop on Speech Synthesis*, 2004.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *Proc. of ICASSP*, 2015.

[5] K. Demuynck, D. Van Compernolle, and H. Van Hamme, “Robust phone lattice decoding,” in *in Proc. ICSLP*, Citeseer, 2006.

[6] G. Saon and J.-T. Chien, “Large-vocabulary continuous speech recognition systems: A look at some recent advances,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 18–33, 2012.

[7] P. C. Woodland and D. Povey, “Large scale discriminative training of hidden Markov models for speech recognition,” *Computer Speech & Language*, vol. 16, no. 1, pp. 25–47, 2002.

[8] N. T. Vu, F. Kraus, and T. Schultz, “Multilingual A-stabil: A new confidence score for multilingual unsupervised training,” in *Proc. of Spoken Language Technology Workshop*, pp. 183–188, 2010.

[9] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech,” in *Proc. of ICASSP*, vol. 2, pp. 554–557, 1993.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, *et al.*, “The Kaldi speech recognition toolkit,” 2011.

[11] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, “Improving deep neural network acoustic models using generalized maxout networks,” in *Proc. of ICASSP*, pp. 215–219, IEEE, 2014.

[12] A. W. Black and P. A. Taylor, “Automatically clustering similar units for unit selection in speech synthesis.,” 1997.

[13] R. Kumar and S. P. Kishore, “Automatic pruning of unit selection speech databases for synthesis without loss of naturalness.,” in *Proc. of Interspeech*, 2004.

[14] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, “Confidence measures for large vocabulary continuous speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[15] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.