신호의 복원된 위상 공간을 이용한 오디오 상황 인지

Le Thanh Hien, 이승룡
유비쿼터스 컴퓨팅 연구실, 컴퓨터 공학과, 경희대학교
e-mail : hien@oslab.khu.ac.kr, sylee@oslab.khu.ac.kr

A new approach technique on Speech-to-Speech Translation

Le Thanh Hien, Sung-young Lee; Young-Koo Lee
Ubiquitous Computing Lab, Computer Engineering Department, Kyung Hee Univerisity, Korea
e-mail : hien@oslab.khu.ac.kr, sylee@oslab.khu.ac.kr;yklee@oslab.khu.ac.kr

Abstraction

We live in a flat world in which globalization fosters communication, travel, and trade among more than 150 countries and thousands of languages. To surmount the barriers among these languages, translation is required; Speech-to-Speech translation will automate the process.

Thanks to recent advances in Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS), one can now utilize a system to translate a speech of source language to a speech of target language and vice versa in affordable manner. The three phase process establishes that the source speech be transcribed into a (set of) text of the source language (ASR) before the source text is translated into the target text (MT). Finally, the target speech is synthesized from the target text (TTS).

1. Introduction and Problems

From the time of RadioRex (1920), the first known ASR system, many new techniques have been conceived to enhance performance. However, most core algorithms used today are based on research from more than 30 years ago. Though researchers wanted to make ASR a science, today it is primarily considered a component of engineering. ASR's performance with languages of limited vocabulary and read speech maybe nearly perfect, but the automatic recognition of speeches of large vocabularies, conversational speaking styles, and different speakers in a large room is far from perfect compared to the human hearing system.

One possible reason for this difference is that, in human hearing, the neural processing system and sound perceiving system work together in a parallel way while, in an ASR system, the level of parallelism is very limited. Speech signals contain much information, including the definition of an acoustic unit. The sequential order of sub-phone and sub-word units, the functioning importance of each language unit, and the hierarchies of information, that needs to be recognized and processed simultaneously. This information is neither perceived nor adequately processed, modeled in current ASR system.

Another problem with ASR concerns the ability to continue to learn. Last year, one ASR system exceeded the performance of one human being on a certain criterion. One wonders, though, if both competitors attempted to learn more and correct their flaws over the last year, whether the ASR system would win again this year. Chances are slim for the ASR considering its limited ability to manage Out-Of-Vocabulary (OOV) words. The competitors would be judged regarding their capacity to figure out whether a word was OOV and, if the system could, whether it could then add that word to its current knowledge base without expensive retraining.

Within the framework of the limitations of ASR as compared to human hearing despite decades of development, we can question current techniques of ASR. We will use two critical measures: The current performance's evaluation criteria and current system's modeling.

2. Better modeling using Maximum Entropy after re-evaluating of ASR

Many techniques have been introduced for ASR, including kernel-based ([2]) and long linear ([3]), but not many are widely supported as current HMM-based systems are. The main reason is that novel techniques may not immediately attain the level of performance of state-of-the-art HMM-based system in terms of WER. It is clearly seen that, if the hunger for better WER is bypassed, we will likely move in a positive direction whereby the modeling of a novel system begin to parallelize recognition.

As an alternative to Byes' decision rule, Maximum Entropy (ME) can be used. The ME principal is modeled completely on what we observe and assumes nothing about what is not observer ([4]). The ME approach can be formulated through the log-linear combination of different feature functions where in the ASR of each feature function can be a dependency (constraint) between the sound and text units. ME have two very nice properties: only one global optimum with convex criterion optimization and algorithms that reach the maximum (e.g. Global Interactive Scaling) and unlimited

feature functions for the current system, which is includes acoustic probability and language model probability, when weighting each feature function with a scaling factor.
The ME-based training method has been successfully applied to the field of natural language processing, such as its application to language modeling ([5]), speech tagging, language understanding ([6]), and statistical machine translation. Unlike natural language processing, ME-based methods have been investigated minimally for automatic speech recognition. ME is employed to estimate the parameters of a direct model for a phoneme recognizer. The approach generalizes the ME Markov Models (MEMM) proposed by such that sequential processes with complex contextual information can be processed.

3. Initial thoughts on improvements to the current SMT

The first and most significant problem regarding SMT is the amount of approximation used in training and decoding. Increasingly powerful learning models have been introduced, but they have to approximate both training and decoding due to the complexities of working with real-world data sets.
The second problem involves decoding. Currently, the decoding algorithms (beam search, A*, integer programming, and optimal decoding) are very dependent on learning models. Thus, the implementation of the search algorithm into the new model is quite expensive.
A third problem coincides with the use of phrase-based translation. Presently, while segmenting a sentence into phrases within a target-language, such as the source-language in a source-channel approach, we assume that the number of segments, K(phrases), is a uniform random number and we segment the sentence into K phrases in what is assumed to be a uniform manner.
Also, with phrase-based translation, a problem exists with phrase re-ordering. While unable to perform a full re-ordering, due to the NP hardness, we should be able to execute a local jump/swap within a short window of phrases.
The final problem centers on the method employed to judge the translation quality. While the current objective and subjective scores are widely used and seem to judge similar systems and gradual changes within one system well. They fail occasionally to compare systems of different approaches, such as SMT and non - SMT
Quantizing discrete probabilities: Vector Quantization is one of the methods by which we can measure Dimensionality Reduction. Different levels of the application of quantization techniques to SMT exist. At first glance, since the order of words in source and target languages are different, quantization may help a model to learn different scales of possible positions of words in a language. This is extremely helpful to those utilizing current learning models and decoding algorithms since the process facilitates the alignment of translation units, words or phrases, in training and re-ordering of those units in decoding. Quantization further helps to reduce the CPU requirement. However, the amount of processing needed to train a statistical model for a huge database exceeds the capacity of existing computers. Thus, though widely used in ASR, quantization remains unpopular in SMT.
Reduced decoding effort: While implementing a decoding algorithm for a new learning model is expensive, this cost can be reduced if one formulates a general search framework. One example involves the use of Weighted Finite State machines whose tools are available as the AT&T FSM toolkit ([6]). In this framework, a search is built as a large Weighted Finite State Transducer (WFST), which encompasses a series of cascaded smaller WFSTs. With such search topology construction, the modification of a WFST component can easily realize the modification or substitution of details in a search algorithm. However, the current implementation of such a search algorithm using the AT&T FSM toolkit is quite slow sin AT&T FSM is not highly optimized for SMT decoding. One improvement would be build one's own FSM dedicated to SMT.
Better evaluation: Multiple metrics should be used more frequently than they are currently utilized. The current usage of metrics may provide a good score to a system that has produced some terrible errors, whether false, ungrammatical, or meaningless, that even a non-professional translator may not commit. We must therefore rely more on multiple reference space in SMT is huge and lacks measures such as edit distance or n-gram matching. Though a new metric could be invented, the inclusion of the usage of multiple metrics would be wiser approach than an attempt to persuade the research community to use a new one would likely prove to be.
More feature functions: Non-contiguous phrases should be utilized together in conjunction with as many feature functions as can be modeled from the translation.

4. Acknowledgement

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2009-(C1090-0902-0002)). Also, it was supported by the IT R&D program of MKE/KEIT, [10032105, Development of Realistic Multiverse Game Engine Technology].

5. References

[1] H.Nanjo and T.kawahara. A new ASR evaluation measure and minimum Bayes-risl decoding for open-domain speech understanding. In Proc. IEEE-ICASSP, Vol.1, pp.1053 { 1056, 2005.
[2] Joseph Keshet, Shai Shalev-Shawartz, Samy Begio, Yoram Singer and Dan Chazan, Discriminative Kernel-Based Phoneme Sequence Recognition, International Conference on Spoken Language Processing (INTER-SPEECH), Pittsburgh, PA,2006
[3] Hong-Kwang Jeff Kyo, Maximum Entropy modeling for speech recognition, ISCLP 2004 Tutorial.
[4] E. Jaynes, Information theory and statistical *mechanics*, Physics Review, Vol.106, no.4, pp. 620-630, 1957.
[5] R. Rosenfeld, A maximum entropy approach to adaptive statistical language modeling, Computer, Speech, and Language, vol.10, no.3, pp. 187228, July 1996.
[6] S. Kumar, Y. Deng, and W. Byrne. 2006. *A weighted finite state transducer translation template model for statistical machine translation*. Journal of Natural Language Engineering.