TOWARD TO RAPIDLY CONFIGURABLE AND CUSTOMIZABLE AUTOMATIC SPEECH RECOGNITION SYSTEM

Nicolae OBJELEAN
Universitatea de Stat din Moldova

Introduction

We propose to resolve two problems of current speech recognizers: 1) the need to record, process, and train for each new vocabulary, and 2) the need to generate a pronunciation dictionary manually. For first problem we propose the use of vacabulary-independent subword unit. We will also investigate the modeling of phonetic variabilities, beyond triphones, using generalized allophones. Other topics in this investigation include acoustic subword units and vocabulary adaptation.

To deal with the second problem, we propose an integrated approach that combines three knowledge sources into an automatic pronunciation dictionary generation system. Since this knowledge sources are relatively independent, we expect the resulting pronunciations to be accurate and consistent.

On of the difficult process in configuring a speech recognizer is the generation of a spelling- to-pronunciation dictionary. This problem is partly alleviated by the use of a very large dictionary and creation of speech language recourses. But in most application, a large dictionary cannotcoverall words in the vocabulary. For example, proper names (such as those for avoice-dialing telephone) and applications-specific terms (medical, legal, and scientific) may be absent. A more serious problem in the addition of new words after a system has been configured. This is problematic because users cannot be expected to have sufficient phonetic knowledge to produce reliable phonetic representations.

1. Problem Statement

Recent receaches [1] has shown that highly accurate speaker-independent speech recognition is possible. These advances eliminated the nee for speaker-dependent trainuing, or the process of using a lage amount of recorded speech to tune system parameters. However, thes speaker- independent system still demand vocabulary-dependent-

training with a large population of speakers for each new vocabulary. They also reguire a manually constracted pronunciation dictionary, which cannot be built or expanded somelevel of phonetic expertise. As speech recognition flourishes and new application emerge, the demand for vocabulary-specific training will become the bottleneck in building speech recognizers, and the difficulty of adding newwords will become the Achiles' heel when customizing speech recognizers. The challenge for our proposed reseatch is to propose, implement, evaluate novel techniques for accurate vocabulary-independent recognition and pronunciation dictionary generation.

Current speech recognition systems are highly vocabulary-dependent, because training on one vocabulary and testing on another leads to a substantial performance degradation. Thus, in order to optimize performance, with current technology, it appears that vocabulary-specific training is necessary. However, vocabulary-specific trainingis an extramely tedious task. For speaker-independent applications, training requires first collecting, checking, and procvessing speech from handreds of speakers. Only then can the actual training, or the process of estimating the parameters of the speech models, commence. This entire process can take several months. In addition to being time consuming, vocabulary-specific training also requires a very large amount of computer time and disk space. Finally, users with new vocabularies must be traned to understand speech collection, data organization, speech processing, and the training process. Another critical aspect of configuring a speech recognizer is pronunciation dictionary generation. Large-vocabulary recognition requires the use of subwords models, so it is necessary to specify the pronunciation of words in terms of subword (usually phonetic) units. The entire dictionary must be prepared before training can commence. Moreover, the addition of every new word requires manual entry of a phonetic representation consistent with the system's existing dictionary. Converting a word into a phonetic units requires knowledge of phonetics, which a typical user lacks. Even with phonetic knowledge this process is time consuming and error-prone.

2. Vocabulary-Independent Models

In order to alleviate the need to train for new vocabularies, we propose to constrict a set of models that are vocabulary-independent. These vocabulary-independent models can be trained once and for all, so that new vocabularies do not require retraining. Since this requires every unit

in the test vocabulary be derivable from the training data, word models are unsuitable. Instead, some subword units must be used.

Good vocabulary-independent units must have two important properties- must be consistent and must be trainable for new tasks and words. Traianbility is important becouse with our hidden Markov speech models [2] require a considerable amount of training data. Optimal performance cam be achieved only after the parameters for all, or nearly all, subword units have been well trained.

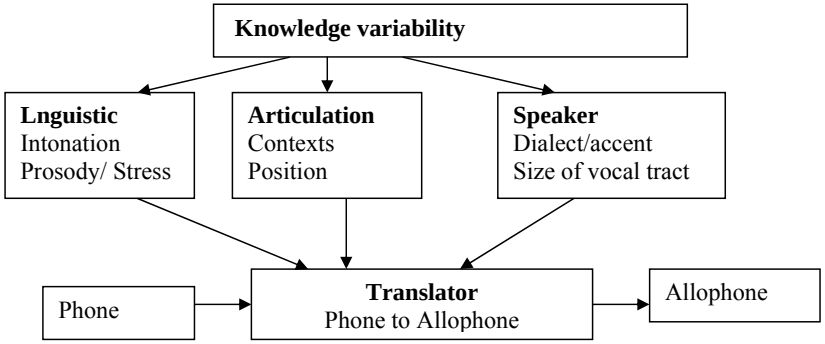
With this in mind, we propose to develop two types ov vocabulary-independent subword models. The first is based on fonetic models, and the second on acoustic models. We will also siscuss techniques that improuve a set of these vocavulary-independent models with vocabulary-adaptation.

3. Phonetic models

Each words may be presented as of subword units most similary phones, and speech production can be considered a process of syntesizing the sequence of phones that correspond to a sentence. Several important continuous spech systems are based on monophones, or simple phone models that do not consider context. The good results of these systems might suggest that monophones are suitable vocabulary-independent units.

These three sources of variability modify various attributes and qualities of each phone, and we call the resultant of this transformation an allophone. The factors affecting the process of transforming a phone to an allophone is illustrated in *Figure 1*.

We propose an architecture in which two knowledge sources generate constraints that guide a phonetic recognizer. In particular, we plan to:



Implement a phonetic recognizer that, given one or two examples of a word, produces the most likely phone sequence, subject to the constraints provided by the other two modules.

Modify a text-to-speech system to produce a network of possible pronunciations for a word given spelling of the word. This network must contain the right pronunciation almost all the time.

Create a probabilistic model that maps spelling to phone sequence. This module will assign probabilities to transitions in the pronunciation network.

4. The Automatic generation systems of pronunciation dictionary

We propose a system that automatically generated the phonetic pronunciation of a new word and given the spelling plus one or two utterance of words. Such a systems could also be used to check e x - isting pronunciation dictionaries. There are three promising strategies for this task:

If phonetic recognition were sufficient accurate, we could simply request the user to speak a new word once, and use the recognized phone sequence as the pronunciation for that world. In practice, however, the best phonetic recognizers [1] are only capable of about 60% recognition accuracy. Yet, the same algorithms and models in these mediocre phonetic recognizers can perform extremely well in word or sentence recognition, because of the additional constraints applied at the lexical and grammatical level. Thus, if sufficient constraints were applied to the phonetic recognizer, accurate recognition would be possible [2].

Is very important in this approach to pronunciation generation is tight integration of various knowledge sources that are relatively independent. [3]. The addition of probabilities to the network further reduces the complexity of the recognition task, and we expect to achieve perfectly recognition and we expect these automatically-generated pronunciations to be more consistent than manually-generated ones, which is especially important to for adding new words. Figure 2 illustrates the System for generation of pronunciation vocabulary and the process of generating the network, assigning the probabilities, as well converting a spoken word and it's spelling into a phone sequence.

The main steps in training and using a pronunciation model [4] for ASR:

1. Obtain a canonical (phonemic) transcription of some training material. A standard pronouncing dictionary is used for this purpose,

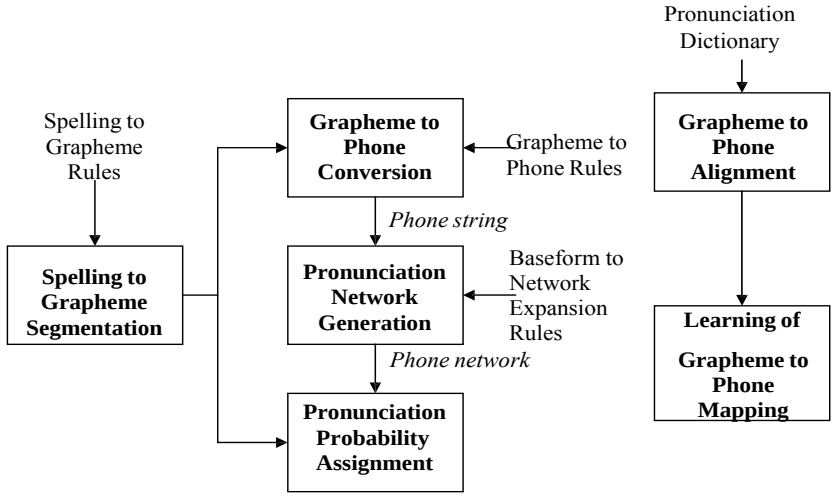


Figure 2. System for generation of pronunciation vocabulary.

with Viterbi alignment when there are multiple pronunciations.

2. Obtain a surface-form (phonetic) transcription of the same material.

3. Align the phonemic and phonetic transcriptions. A dynamic programming procedure based onphonetic feature distances is used for this purpose.

4. Estimate a decision-tree pronunciation model. A decision tree is constructed to predict the surface form of each phoneme by asking questions about its phonemic context.

5. Perform recognition with this pronunciation model. A dictionary-based phoneme level recognition network is obtained from a word lattice generated by an initial recognition pass.

6. Full training set retranscription. Starting with the canonical transcription of the entire acoustic training set, the pronunciation model is used to create pronunciation networks represent.

We propose the architecture of systems in which two knowledge [5] sources that guide a phonetic recognizer.

•Using of phonetics information for produces the most likely phone sequence.

• Modify a text-to-speech system to produce a network of possi-

ble pronunciations for a word given spelling of the word for creation of speech synthesizes systems. This network must contain the right pronunciation of all words in the time.

• Creating a probabilistic model that maps spelling to phone sequence. This module will assign probabilities to transitions in the pronunciation network.

We expect that we will have an order of magnitude more training data within the thirst year of the proposed research.

This will substantially improve the accuracy. Moreover, we will modify the phonetic recognizer to use generalized triphones and eneralized allophones. Finally, isolated words will be used for pronunciation generation, which makes the recognition task easier, since the speech will be much more carefully spoken. In finale we can to construct a language knowledge and phone speech resources by using this system, and by using new subword models. We believe that system aided creates new probabilistic state grammar and can accurately map words into corresponding phone sequence.

Conclusions

This approach proved phonetically word analysis of the text, which examines the structure of each words and aiding the generation of vocabularies for new tasks and give the language knowledge. Also the software developed for collection of language resources, creation phonetically models of the words and using its for generation pronunciations dictionary.

References

[1] Trunin-Donskoi, V.N. The Automatic recognition of the alphabet as a means of comparison of technical opportunities of speech recognition systems. Problems of elaboration the speech understanding systems. Moscow, Science, p. 30-46, 1980.

[2] Lee, K.F.,A connectionist Approach to Continuous Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing. April, 1998.

[3] Objelean ,N. Developing Tools For Colection Language Resources And Building New Voice Web Application The 17th Conference On Applied And Industrial Mathematics Caim Constanta, Romania, September 17-20, 2009, http://www.anmb.ro/ro/conferinte/caim2009/abstracts.pdf

[4] Objelean N. Modeling Pronunciation Dictionary for Adaptation vocabulary of Speech recognition System. Pp. 282-283.Conferenta jubiliara tehnico-stiintifica a colaboratorilor. UTM. Chisinau 2006.

[5] Objelean N., Building Systems for Rapidly Analysis and Modelling of Speech for Creation of Language Resources, Proceedings of 7th International Conference on DAAS, 27-29, 2004. Suceava Romania, p.472-474.