Statistical and Linguistic Knowledge Based Speech Recognition System: Language Acquisition Device for Machines

Challa Sushmita, Challa Nagasai Vijayshri and Krishnaveer Abhishek Challa

Abstract Today's speech recognizers use very little knowledge of what language really is. They treat a sentence as if it would be generated by a random process and pay little or no attention to its linguistic structure. If recognizers knew about the rules of grammar, they would potentially make less recognition errors. Highly linguistically motivated grammars that are able to capture the deeper structure of language have evolved from the natural language processing community during the last few years. However, the speech recognition community mainly applies models which disregard that structure or applies very coarse probabilistic grammars. This paper aims at bridging the gap between statistical language models and elaborate linguistic grammars. Firstly an analysis of the need to integrate the conventional Statistical Language Models with the modern Linguistic Knowledge based language models is made, thereby justifying the Statistical and Linguistic Knowledge based Speech Recognition System which is asymptotically error free.

Keywords Statistical language models · Speech recognizer · Linguistic knowledge · Speech recognition

1 Introduction

The aim of automatic speech recognition is to enable a machine to recognize what a human speaker said. A machine that can "hear" can be helpful in many ways. The user can control the machine by voice, which keeps his hands and eyes free for

C. Sushmita (✉) · C.N. Vijayshri
Andhra University, Visakhapatnam, India
e-mail: challa.abstract@gmail.com

C.N. Vijayshri
e-mail: shri.vijayshri@gmail.com

K.A. Challa
Gayatri Vidya Parishad, Visakhapatnam, India
e-mail: com2mass@gmail.com

S.C. Satapathy et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 434,
DOI 10.1007/978-81-322-2752-6_60

other tasks, it can save the user from typing vast amounts of text by simply dictating it, the recognized speech can be used to index speech such as broadcast news which allows efficient document retrieval, or the system may even understand what the user intends to do or answer his questions. These examples illustrate that speech recognition is an important aspect of improving human-machine interfaces and thus making machines more usable and user friendly.

It was believed, that as soon as the spectrum of a speech signal could be computed fast enough, the speech recognition problem could be easily solved. Although thousands of researchers around the world worked on the problem for more than half a century, the task must be still considered to be unsolved. In difficult acoustical environments machines perform orders of magnitude worse than humans.

How was such a misinterpretation possible? On one hand the speech recognition problem is often largely underestimated because it is so natural for human beings to listen to others and understand them. We are not aware of the tremendous amount of variability present in a speech signal. We can understand people we never met before, we are able to recognize a huge amount of different words in continuous speech, and we are even able to understand ungrammatical utterances or expressions we have never heard before. We are able to perform so well because we include a wide variety of knowledge sources: we have prior knowledge about the syntax and semantics of a language, we can derive the meaning of new words by analogy, we use situational clues like the course of a dialogue and we have access to all experiences we made in our live and all knowledge about the world we have. Machines cannot keep up with that.

Written language consists of a sequence of discrete symbols, the letters of the alphabet. These symbols are uniquely identifiable and do not interact. The boundaries of a word are well defined as words are separated by spaces. This is still true for the smallest linguistic elements of speech, the phonemes. In written form, these are discrete symbols as well. However, the situation changes dramatically when we are going from written form to spoken form, or more specifically if we look at a speech signal.

A speech signal contains a tremendous amount of variability from several sources. There is no one-to-one relationship between letters or phonemes and their physical realisation in a speech signal:

- The acoustic realisation of a phone largely depends on the individual speaker properties such as sex, vocal tract shape, origin, dialect tone coloration, speaking rate, speaking style (normal, whispering, shouting), mood and health.
- The pronunciation of a particular phone is influenced by its phonetic context (coarticulation). This influence may span several phones and even syllable and word boundaries.
- Allophonic variants and phoneme variations.
- The signal is altered by the room characteristics like reverberation, the microphone characteristics, signal coding and compression, as well as background noise.

In order to convert speech to text, a description of the acoustic events of speech alone is not sufficient. To resolve ambiguities, knowledge about the language at hand is indispensable and plays a very important role in speech recognition.

2 Language Models for Speech Recognition

A language model (LM) is a collection of prior knowledge about a language. This knowledge is independent of an utterance to be recognized. It therefore represents previous knowledge about language and the expectations at utterances. Knowledge about a language can be expressed in terms of which words or word sequences are possible or how frequently they occur.

Language models can be divided into two groups. The criterion is whether the model is data driven or expert-driven:

- **Statistical language models**: If the model is based on counting events in a large text corpus, for example how frequent a certain word or word sequence occurs, the model is called to be a statistical language model. Such a model describes language as if utterances were generated by a random process. It is therefore also known as stochastic language model [6, 7].
- **Knowledge based models**: If the knowledge comes from a human expert the model is called knowledge-based language model. Such linguistic knowledge could for example include syntax, the conjugation of verbs or the declension of adjectives. The basis of such a model does not rely on counting observable events, but rather the understanding of the mechanisms, coherences and regularities of a language. If this knowledge is defined by rules, such models are also called rule-based models [2].

Since statistical language models are the most commonly used models, they will be discussed first. Consequently, there is a description of the key idea, the advantages and the limitations of statistical LMs. The limitations will motivate the use of knowledge based models and the approach that was taken in this thesis.

2.1 Statistical Language Models

A statistical LM aims at providing an estimate of the probability distribution P(W) over all word sequences W. It must be able to assign a probability to each possible utterance. The conditional probabilities must be estimated on large amounts of texts related to the recognition task at hand. The number of frequencies that must be counted and stored for this model is prohibitive. The longer the conditioning history gets, more and more strings will never occur in the training data. An obvious solution is to limit the length of the histories by assuming that the probability of

each word does not depend on all previous words, but only on the last N − 1 words which leads to the so called N-gram language model.

An N-gram language model assumes that the probability of a word is not influenced by words too far in the past. It considers two histories to be equivalent, if they have their last N − 1 words in common. With decreasing N the approximation gets coarser and the space requirements decrease. The N-gram is currently the most widely used language model in speech recognition [1, 2].

The simplicity of the model, its easy integration into the decoding process and its ability, at least to some extent, to take semantics into account, contribute to its success. It is also attractive because it is completely data driven, which allows engineers to apply it without requiring detailed knowledge about the language at hand.

However, despite of its success, the word N-gram language model has several flaws:

- **False conditional independence assumption**: The N-gram model assumes that a word is only influenced by its N − 1 preceding words and that it is independent from other words farther in the past. It assumes that language is generated by a Markov process of order N − 1, which is obviously not true.
- **Saturation**: The quality of N-gram models increased with larger amounts of data becoming available online. However the improvement is limited due to saturation. Bigram models saturate within several hundred million words, and trigrams are expected to saturate within a few billion words.
- **Lack of extensibility**: Given an N-gram model it is difficult or even impossible to derive a new model which has additional words. The information contained in the model is not helpful to derive N-grams containing new words. Grammars, on the other hand, are able to generalize better because they are based on the underlying linguistic regularities.
- **Lack of generalization across domains**: N-grams are sensitive to differences in style, topic or genre between training and test data. The quality of an N-gram model trained on one text source can degrade considerably when applied to another text source, even if the two sources are very similar.

N-grams fail on constructions like in the following example sentence:

"The dogs chasing the cat bark."

The trigram probability P(bark/the cat) will be very low because on one hand cats seldom bark, and on the other hand because a plural verb (bark) is unexpected after a singular noun (cat). Nevertheless this sentence is completely sound. The verb (bark) must agree in number with the noun (dogs) which is the head of the preceding noun phrase, and not with the noun that linearly precedes it [9].

2.2 Knowledge Based Language Models

Undoubtedly, written language and spoken language follow certain rules such as spelling and grammar. In a knowledge-based approach these rules are collected by experts (linguists) and are represented as a hand-crafted formal system. This system allows deciding if a sentence belongs to the language defined by the rules, and if that is the case, to derive its syntactic structure. The knowledge is explicitly available [4].

In contrast to a statistical LM, no training data is needed for a knowledge-based system. This can be advantageous if no or only a small amount of (annotated) data is available from a specific domain. At the same time this means that the lexicon can be easily extended.

The knowledge based approach faces several problems. One is of course the difficulty to build a formal system which appropriately reflects the phenomena of a natural language. The main problem that a speech recognizer has to deal with is the binary nature of a qualitative language model. If no appropriate measures are taken, the system is only capable of recognizing intra-grammatical utterances. This is quite a strong limitation, since a recognizer should be able to transcribe extra-grammatical utterances as well. The lack of frequencies of a purely rule-based system is disadvantageous if the recognizer has several hypotheses to choose from which are all intra-grammatical. For example, the sentences "How to recognize speech?" and "How to wreck a nice beach?" are both syntactically correct, however the first is a priori more likely and should be preferred.

Thereby there is a need to integrate the conventional Statistical Language Models with the modern Knowledge based language models leading to a Statistical and Linguistic Knowledge based Speech Recognition System which is asymptotically error free [3, 8].

3 Statistical and Linguistic Knowledge Based Speech Recognition System

3.1 N-Grams Derived from a Statistical Grammar

Stochastic grammars have the advantage that they typically have much less parameters than an N-gram model. Stochastic grammars can thus be more reliably estimated from sparse data than N-grams. However, N-grams can be more easily integrated into the decoder without requiring a parser. The idea is therefore to combine the advantage of reliably estimating the parameters of a stochastic grammar with the ease of integration of N-gram models. This is accomplished by estimating N-gram probabilities from a stochastic grammar instead of using the N-gram counts of sparse data. Including natural-language constraints into the decoder can be desirable for two reasons: First, decoding can be more efficient due

to the reduced search space, and second, it may improve recognition accuracy. The advantage is that undesired, extra-grammatical sentences can be ruled-out early and that low scored intra-grammatical sentences can be saved from being pruned away. To include a grammar into a Viterbi decoder it must be possible to process the grammar left-to-right as the Viterbi-algorithm runs time-synchronously [1–3, 8].

If the grammar is regular, it can be modelled by a finite state automaton and directly integrated into the recognition network of an HMM recognizer; Some natural language phenomena cannot be described in terms of regular grammars or are more elegantly formulated by a context-free grammar. It is not feasible to compile CFGs into a static, finite state transition network because the number of states could be unmanageably large or infinite [5].

However, due to pruning only a part of the state transition network is active at each point in time, therefore a CFG can be realized as a network by dynamically extending the necessary part of the finite state network.

The system incrementally extends the recognition network of a Viterbi decoder by a NL parser and a unification-based CFG. The recognition network is generated on the fly, by expanding the state transitions of an ending word into all words which can follow according to the grammar. It does so by predicting terminal symbols in a top-down manner; non-terminal symbols on the right-hand-side of context-free rules are expanded until a terminal is found.

The dynamic approach was extended by a probabilistic component. It uses a SCFG to compute a follow set and word transition probabilities for a given prefix string. If the prefix string is parsable the SCFG is used to compute the probability distribution of possible following words. If the string cannot be parsed, the system falls back to bigram probabilities instead.

The idea behind predict and verify is very similar to the dynamic generation of partial grammar networks. The main difference is that in the dynamic generation approach the parser is driven by the HMM decoder, while in the predict and verify approach the emphasis is put on the parser which drives the recognition process. It is based on predicting the next word or the next phone in a top down manner and is also called analysis by synthesis. A word or a phone is assumed to be present if its maximal likelihood over all possible ending points is larger than a threshold [3, 5].

References

1. Brown P.F., Della Pietra V.J, deSouza P.V., Lai J.C., and Mercer R.L.: Class-based n-gram models of natural language. Computational Linguistics, 18(4):467–479 (1992).
2. Brill E., Florian R., Henderson J., and Mangu L.: Beyond N-grams: Can linguistic sophistication improve language modeling? In COLING/ACL 1998, pages 186–190, Montreal, Canada (1998).
3. Beutler R., Kaufmann T., and Pfister B.: Integrating a non-probabilistic grammar into large vocabulary continuous speech recognition. In Proceedings of the IEEE ASRU 2005 Workshop, pages 104–109, San Juan (Puerto Rico), (2005).

4. Beutler R., Kaufmann T., and Pfister B.: Using rulebased knowledge to improve LVCSR. In Proceedings of ICASSP, pages 829–832, Philadelphia (2005).
5. Burshtein D.: Robust parametric modeling of durations in hidden Markov models. In Proc. of ICASSP, volume 1, pages 548–551, Detroit, Michigan U.S.A (1995).
6. Charniak E.: Statistical parsing with a contextfree grammar and word statistics. In Proceedings of AAAI/IAAI, pages 598–603 (1997).
7. Gillick L. and Cox S.: Some statistical issues in the comparison of speech recognition algorithms. In ICASSP, pages 532–535 (1989).
8. Harper M., Jamieson L., Mitchell C., Ying, S. Potisuk G., Srinivasan P., Chen R., Zoltowski C., McPheters L., Pellom B., and Helzerman R.: Integrating language models with speech recognition. In "Proceedings of the AAAI- Workshop on Integration of Natural Language and Speech Processing" (1994).
9. Rosenfeld R.: Two decades of statistical language modeling: Where do we go from here? Proceedings of the IEEE, 88(8):1270–1278 (2000).