Dissertation ETH No. 19507

Speech Recognition Techniques for Languages with Limited Linguistic Resources

A dissertation submitted to the
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by
MICHAEL GERBER
Dipl. El.-Ing. ETH
born December 14, 1975
citizen of Langnau i. E. (BE), Switzerland

accepted on the recommendation of
Prof. Dr. Lothar Thiele, examiner
Prof. Dr. Jean-Pierre Martens, co-examiner
Dr. Beat Pfister, co-examiner

2011

Abstract

There are several thousand languages in the world and each language has a multitude of dialects. State-of-the-art speech recognition techniques, which are usually based on transcriptions, are however only available for a few languages because of the lack of acoustic and textual resources which are necessary to build these recognizers.

In this thesis we aim at the development of speech recognition technologies for languages with limited or no resources. For many applications such as the control of machines or home appliances by voice it is not necessary to have a continuous speech recognizer with a large vocabulary. It is then possible to resort to techniques which need only very little language-specific resources.

In order to build isolated word recognizers for any language we relied on speech recognition techniques with an utterance-based vocabulary. In these techniques each word of the vocabulary is defined by one or several sample utterances. This way of defining the vocabulary is language-independent and has the further advantage that it can be done by everybody since no expert knowledge is required.

To improve the recognition rate of speech recognition with an utterance-based vocabulary we worked with two techniques: the first one based on dynamic time warping in combination with specially trained artificial neural networks and the second one based on hidden Markov models with data-driven sub-word units.

With the availability of moderate resources from the target language we were able to develop a recognizer technique which yielded

comparable results to a transcription-based recognizer which requires in contrast to our technique a pronunciation dictionary to build the word models. When no resources of the target language were available and resources from other languages than the target language had to be used instead, the performance of transcription-based recognition was not achievable with the utterance-based recognizer techniques developed in this thesis. Yet, in this case the developed approaches allowed to halve the error rate of isolated word recognition with an utterance-based vocabulary compared to a standard approach based on dynamic time warping using the Euclidean distance measure.

We also applied the developed techniques to other applications such as acoustic data mining. In this way it was possible to tackle these problems for speech signals of any language since the developed techniques do not require resources of the target language.

Kurzfassung

Weltweit existieren einige Tausend Sprachen, und in jeder Sprache werden viele verschiedene Dialekte gesprochen. Spracherkenner, welche dem Stand der Technik entsprechen, stehen allerdings nur in den wenigsten Sprachen zur Verfügung, da zu ihrer Implementierung umfangreiche akustische und linguistische Ressourcen notwendig sind.

In dieser Arbeit haben wir Techniken entwickelt und getestet, welche die Spracherkennung in Sprachen mit wenigen oder keinen Ressourcen verbessern. Für viele Anwendungen, wie zum Beispiel die Steuerung von Maschinen oder Haushaltsgeräten, ist es nicht nötig, einen kontinuierlichen Spracherkenner mit einem grossen Vokabular zur Verfügung zu stellen. Mit diesen geänderten Anforderungen werden Techniken, welche keine sprachspezifischen Ressourcen benötigen, möglich.

Um die Erkennung von isolierten Wörtern in beliebigen Sprachen zu ermöglichen, haben wir Techniken, die ein Vokabular verwenden, das auf Musteräusserungen basiert, verbessert. Bei diesen Techniken wird jedes zu erkennende Wort durch eine oder mehrere Musteräusserungen definiert. Neben der Sprachunabhängigkeit haben diese Techniken auch den Vorteil, dass ein Vokabular von jedermann definiert werden kann, da kein Expertenwissen nötig ist.

Zur Verbesserung musterbasierter Spracherkennung haben wir grob mit zwei Techniken gearbeitet: die erste basiert auf dynamischer Zeitanpassung in Kombination mit speziell trainierten künstlichen neuronalen Netzen, und die zweite basiert auf *Hidden-Markov-Modellen* mit speziellen akustisch motivierten Sprachelementen.

Wenn einige wenige Ressourcen der Zielsprache zur Verfügung standen, konnten wir mit den entwickelten Techniken Erkennungsraten erreichen, welche jenen eines dem Stand der Technik entsprechenden, Aussprachewörterbuch-basierten Erkenners in nichts nachstehen, auch wenn dieser mehr Ressourcen wie zum Beispiel ein Aussprachewörterbuch benötigt. Falls gar keine Ressourcen in der Zielsprache zur Verfügung standen und auf Ressourcen einer anderen Sprache für das Training der Modelle zurückgegriffen werden musste, konnten die Erkennungsraten von Aussprachewörterbuch-basierten Erkennern nicht erreicht werden. Die Fehlerraten welche wir mit unseren Erkennern erreichten, waren allerdings trotzdem nur halb so gross wie jene von konventionellen Mustervergleich-Erkennern.

Wir haben die entwickelten Techniken auch für andere Anwendungen, wie zum Beispiel die Suche von lautlich ähnlichen Abschnitten, wie Wörtern, in zwei Sprachsignalen angewendet. Diese Anwendungen werden dank den neuen Techniken in beliebigen Sprachen möglich.