

Preface

Automatic speech recognition systems are increasingly applied for modern communication. One example are call centers, where speech recognition based systems provide information or help sorting customer queries in order to forward them to the according experts. The big advantage of those systems is that the computers can be online 24 h a day to process calls and that they are cheap once installed.

Travelers can find speech-to-speech translation systems commercially available, either via cellular phone or based on hand-held devices such as PDAs. These translation systems typically provide translation from English to Japanese (or in case of American military systems, to Arabic) and back. Pressing a push-to-activate button starts the recording of an utterance, which is recognized and translated. The translated sentence is then played to the communication partner with a text-to-speech (TTS) module.

Speech control is also a common feature of car navigation devices, for command and control purposes as well as destination input. Undeniably, speech control increases comfort. It is also undisputed that speaking a telephone number whilst driving rather than typing it on the tiny cellular phone keyboard is an important safety measure, as the driver can still look and concentrate on traffic without too much distraction by the input procedure. Speech controlled music players will enter the mass market soon.

Moreover, in many of these applications of speech recognition, there are cases when the speaker is not expressing himself in his mother tongue. Customers of call centers are people who are not necessarily citizens of the country where the service is offered nor have they mastered the local language to a native level of proficiency. Translation systems for travelers as sold e.g. in Japan assume that most foreign dialog partners the owner of a Japanese–English translation system encounters speak English, native or as a second language. Car navigation systems must offer drivers the possibility to set a destination abroad, therefore good systems have to support speech input of foreign place names. Finally, given the strong internationalization of the music market, speech controlled music players must cope with non-native speech input, e.g. for English language songs in Germany or Spanish language songs in the United States.

A speech recognition system recognizes words as a sequence of phonemes defined in a pronunciation dictionary. These sequences do not entirely match non-native speaker utterances as they deviate from the standard pronunciations of words, inserting and omitting sounds as typical for the phonetic contexts of the native language. They especially generate different sounds that are more familiar from the speakers mother tongue but do not fully match the phonetic inventory of the language the speaker has not fully mastered. For both humans and machines, these deviations are a big hurdle to understand what a non-native speaker says. By listening to accented speech for some time, humans can learn the specific accent patterns and adapt to them to some extent.

The target of this research is to provide a method that adjusts an automatic speech recognition system so that it can recover some of the errors caused by non-native pronunciation. We relax the pronunciation dictionary constraints for recognition of non-native speech. Then by training on a non-native speech sample, we factor in the specific pronunciation error patterns of each accent without attempting to represent them explicitly.

The authors would like to thank Seiichi Yamamoto for providing the opportunity to do research at ATR. Furthermore we thank Konstantin Markov for valuable discussions and helpful suggestions and Nobuaki Minematsu and Frank Soong for analyzing this work and asking questions.

The first author is especially grateful to Heinrich Niemann, Elmar Nöth, Yoshinori Sagisaka and Harald Singer for introducing him to the marvels of automatic speech recognition and its practical application.

We acknowledge the valuable insights the research of Norbert Binder, Tobias Cincarek, Bi Lige, Martin Raab and Wenzel Svoyanovsky have contributed to this work.

Thanks go to Nick Campbell and JST/CREST ESP for granting access to the expressive speech database.

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology (NICT) of Japan entitled, “A study of speech dialog translation technology based on a large corpus”.

Statistical Pronunciation Modeling for Non-Native
Speech Processing

Gruhn, R.E.; Minker, W.; Nakamura, S.

2011, X, 114 p., Hardcover

ISBN: 978-3-642-19585-3