MULTI-LINGUISTIC SPEECH RECOGNITION

Cătălin Chivu

“Transilvania” University of Brasov, Romania, Department of Economic Engineering and
Manufacturing Systems, Str. Parcul Mic, nr.13, bl.110, ap.31, Brasov, Romania,
e-mail: chivuc@acasa.ro

Abstract: The present paper proposes a utilization method of speech recognitions engines either by creating personalized grammars for a specific language or by including the recognition engine in personalized programs where the programmer defines a set of words that will be recognize in the desired language.

Keywords: speech recognition, recognition engine, phoneme, grammar

1. INTRODUCTION

Many companies produce speech recognition engines that are quite performing but, generally speaking, these recognition engines are part of oriented programs that recognize the language of the producers (the most of them being in English and a few in other languages as Germany, French, Japanese or Chinese).

The present paper proposes a utilization method of speech recognitions engines either by creating personalized grammars for a specific language or by including the recognition engine in personalized programs where the programmer defines a set of words that will be recognize in the desired language.

First of all, a recognition engine to be able to be personalized (to recognize words from specific languages) must have defined the following elements:

- numbers of phoneme and the code associated with these phonemes; these correspond to the specific language;
- the links between phonemes;
- the set of words that will be recognize and that belongs to the specific language;
- at least one elementary grammar, which defines a relation between the words that are part of the set that will be recognized.

The words included in the set that will be recognized should be defined based on the phonemes specific to that language.

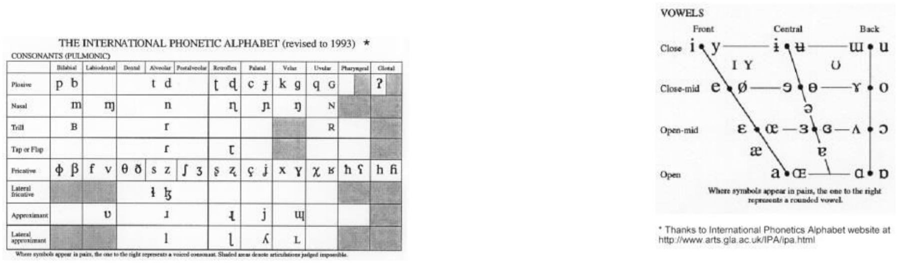
The present paper concentrates upon the speech recognition engine designed by Microsoft. The native language of this engine is, of course, American English. This recognition engines has remarkable performances if it is used for the native language. As it was said before, the difference between the two recognition engines designed for specific languages is not necessary a structural one but can one given by the difference between the set of phonemes, set of words from vocabulary and the grammars defined for this vocabulary.

Thus, in the following will be analyzed these three element for the particular case of the Microsoft speech recognition engine.

2. MULTILINGUAL DATA RESOURCES

What kind of resources are available for multi-lingual ASR :

-Speech and Text Resources



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993) *

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap				ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

* Thanks to International Phonetics Alphabet website at http://www.arts.gla.ac.uk/IPA/ipa.html

Fig.1. Multi-language speech recognition - phonetics

-European Language Resources Association -ELRA www.elra.info

-Linguistic Data Consortium -LDC www.ldc.upenn.edu

-SpeechDat family www.speechdat.org

-Including SpeechDat, SpeechDat II, Speecon, SALA, SpeechDatCar, etc.

-Databases for tens of languages available for public

-Databases tend to be expensive: E.g. American English SpeechDatCar database: 120k€ for commercial rights; E.g. French SpeechDatCar database: 182k€ for commercial rights

3. PRONUNCIATION MODELING

Pronunciation modeling statistical methods implies:

-All require variable amount of training data –depending on the irregularity of the language

-N-gram modeling applicable for semi-regular languages require quite large memory to store all N-gram rules especially when N >= 3

-Neural network can learn complex irregularities and has pretty good rule generalization and fast decoding.

-Decision tree, can learn any non-overlapping exceptions, can combine several rules to reduce memory consumption, complex decoding process

4. INTERNATIONAL PHONEME REPRESENTATION

It can be created pronunciations for words that are not currently in the lexicon using the phonemes represented in the attached appendices. The proposed phoneme set is composed of a symbolic phonetic representation (SYM).

The SYM representation can be entered to create the pronunciation by using the XML PRON tag, or by creating a new lexicon entry. Each phoneme should be space delimited.

The engine is passed a USHORT structure called SPPHONEID (a number between 1 and n where n is the total number of phonemes for that language). The conversion from the SYM to SPPHONEID occurs in the SAPI PhoneConverter.

SAPI-compliant engines are required to accept the PHONEID representation, and produce an articulation. The specific allophonic articulation is defined by the engine. There is no provision for support of phonemes outside the SAPI phoneme set.

4.1. Main goals for defining the language dependent phoneme set:

- Provide an engine-independent architecture for application developers to create user and application lexicons.
- Make the English phonetic table simple enough to be used and understood by non-linguists who use the American English phoneme set.

4.2. International phoneme use

Using the international phoneme schema, can be created a phoneme set which can be used for each language independently. Using the numeric representation as opposed to the International Phonetic Alphabet (IPA) code will eliminate some of the problems regarding the possible differences in the IPA values for the same phonemes. Hence, an 'r' in English will correspond to a certain number (38) and an 'r' in French may correspond to a different number. It is up to the individual engine to provide the exact IPA value for the two 'r's.

Each language will be associated with a set of phonemes numbered from 1 to X. You can use either the symbolic representation or the number representation to enter the pronunciation. Since you are probably not a linguist, the IPA code will probably have little meaning.

Please note that consistent pronunciation is NOT a goal, while predictable pronunciation is. Using the phoneme set, an application developer can guarantee a minimal pronunciation, but not the exact allophonic expression. So, the word "first" will always be pronounced as "first", never as "fist", "feast", etc, but the accent may be slightly different due to the fact that the internal allophone values may differ.

5. CONCLUSION

Creating an application, based on the Microsoft speech recognition engine, which is able to understand words and phrases in other languages than those for it was designed, it is possible. The performances of speech recognition engine will be a little lower because there is no possibility to re-train the system thus it will be able to adapt itself to the pronunciations significantly different for different languages.

REFERENCES

[1] Boulard, H., D'hoore, B., Boite, J.-M. (1994) Optimizing recognition and rejection performance in word spotting systems, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I-373 -I-376, Australia.
[2] Hariharan, R., Häkkinen, J., Laurila, K. (2001) Robust end-of-utterance detection for real-time speech recognition applications, ICASSP 2001, Salt Lake City.
[3] Lamel, L., Rabiner, R., Rosenberg, J., Wilpon, J. (1981) An improved end-point detector for isolated word recognition, IEEE ASSP magazine, pp. 777-785, 1981.
[4] Suontausta, J. Tian, J. (2003) Low memory decision tree method for text-to-phoneme mapping, IEEE ASRU workshop, USA
[5] Viikki, O. (2001) ASR in Portable Wireless Devices, Proc. IEEE Automatic Speech Recognition & Understanding, Madonna diCampiglio, Italy.