
SPEECH PROCESSING

Abstract

The mission of the speech processing group is to advance text-to-speech synthesis, speech recognition, and speaker verification, both in research and in technology development. The relation between speech and text is in the focus of our work. We pursue mainly interdisciplinary approaches that require competence in fields as diverse as digital signal processing, statistical modeling and computational linguistics (e.g. grammar formalisms and parsers for natural languages). The extensive

application of rule-based linguistic knowledge in morphology and syntax is precisely what distinguishes our work from the work of most other speech research groups. Several ongoing projects contribute to our long-term aims. Three of them, namely polySVOX, RULAMO and SpVeri are outlined in the sequel.

Section

Computer Engineering

Head

Prof. Dr. Lothar Thiele

Project Coordinator

Dr. Beat Pfister

TIK Participants

Thomas Ewender

Michael Gerber

Sarah Hoffmann

Tobias Kaufmann

Harald Romsdorfer

Web

www.tik.ee.ethz.ch/~spr/Proj_SPG.html

SPEECH PROCESSING

polySVOX: Speech synthesis from mixed-lingual texts

Since many years there is an increasing tendency to mix languages, particularly in multilingual countries like Switzerland. Mixing means using foreign names and terms, but particularly also including items of various size from other languages. Sentences like «Das Server Management inklusive SAP Applikationsmanagement wurde bei uns outsourct.» are quite common.

In order to produce a natural pronunciation for such texts, a text-to-speech synthesizer needs to identify foreign inclusions, no matter whether they are as short as a part of a word or as long as a whole sentence. This is particularly difficult in the case of interlingual homographs as shown by the example in Figure 1.

We have developed a text analysis component for the languages German, French, Italian and English with very high language detection accuracy (see [1]). A preliminary speech production component allows to produce synthetic speech for these four languages with a single voice (currently with German prosody only). These two components are essential parts of our polySVOX system. It is the first text-to-speech system world-wide which applies morphological and syntactic analysis to deal with mixed-lingual input text and generates mixed-lingual speech. Example sentences can be found on our webpage.

A major challenge in text-to-speech synthesis in general and in the case of mixed-lingual texts in particular is the generation of appropriate prosody (speech melody and rhythm). The aim of two recently started complementary projects is to investigate the prosody of speech from professional speakers and to implement this knowledge by means of statistical models in our text-to-speech system.

In these projects, we mainly cooperate with SVOX AG, a leading edge text-to-speech synthesis technology provider. Furthermore, there is an ongoing informal collaboration between several partners of the former COST Action 258. The projects are supported by CTI and ETH.

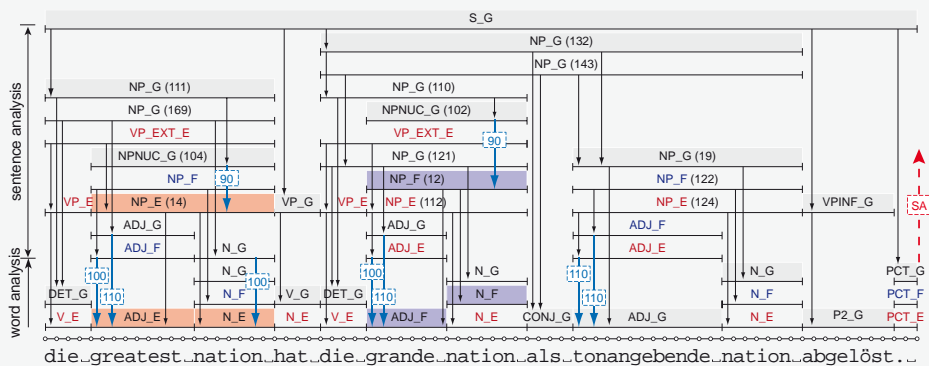


Fig. 1: Representation of the result from morphological and syntactic analysis of the mixed-lingual German sentence: «Die Greatest Nation hat die Grande Nation als tonangebende Nation abgelöst.»

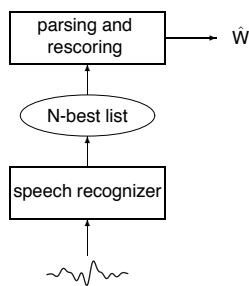


Fig. 2: A standard speech recognizer is used to detect the N best hypotheses from the speech signal. These word sequences are processed by the parser and rescored taking the degree of grammatical correctness into account.

RULAMO: Rule-based language model for speech recognition

State-of-the-art speech recognizers use a statistical language model for compensating the weaknesses of the HMM-based acoustic model. The most frequently used language models are word sequence statistics such as N-grams. The main reasons for the success of the N-grams are their simplicity and that they can easily be integrated with the acoustic models which allows for efficient processing.

On the other hand, N-grams clearly model natural language on a very superficial level, mostly ignoring the dependencies between non-adjacent words (e.g. non-local dependencies, valency and agreement). This is particularly problematic for highly inflected languages with relatively free word order, such as German.

To compensate for these weaknesses, we plead to apply in addition to N-grams a rule-based language model (consisting roughly of a lexicon, a grammar and a parser) that checks the syntactical correctness of the recognized word sequences. The idea is to decrease the word error rate by favoring syntactically correct phrases over incorrect ones in a second processing stage as shown in figure 2. To decide on syntactical correctness is a very difficult problem, however. Recent progress in computational linguistics, mainly in syntax theories and grammar formalisms, suggests feasibility.

So far we have developed such a rule-based language model. In a rather simple dictation task we could significantly reduce the recognition error rate with this language model (see [3] and [4]). In order to give stronger evidence for the benefit in general, it has to be tested in a much more challenging scenario such as broadcast news transcription. For that purpose the coverage of the

linguistic components (lexicon and grammar) and the capability of the parser to process long sentences needs to be drastically increased (cf. [5]).

The ultimate target of the project is to show that the best statistical speech recognizers still can be improved by our rule-based language model. This is very challenging and requires collaboration on the international level. In particular we are collaborating with the LIMSI spoken language processing group, one of the key players in German speech recognition.

The project is supported mainly by the SNF, partly by the NCCR IM2.

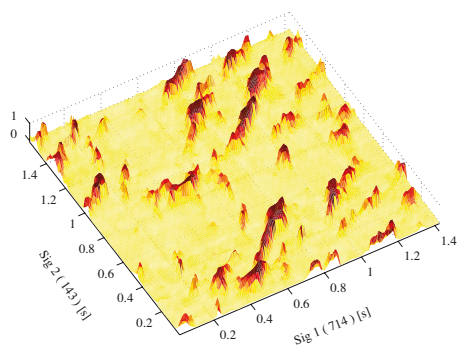


Fig. 3: Probability matrix spanned by the German numbers 714 and 143 uttered by two different speakers. It shows the probability that the phonemes at time x in signal 1 and at time y in signal 2 are the same. Common segments show as ridges, e.g., the word «hundert» which occurs in both numbers shows as a ridge between 0.4 and 0.8 s on x-axis and between 0.2 and 0.6 s on the y-axis.

SpVeri: Speaker verification

In some applications it is necessary to know whether two given speech signals are from the same person or not. Since the two speech signals generally have different wordings, a text-independent method is required. The state-of-the-art for text-independent speaker verification is based on Gaussian mixture models (GMM). These models neglect the sequential order of the phonemes in the speech signal. We argue for an approach which makes use of this order, namely pattern matching. Pattern matching is a text-dependent method, i.e., it can be applied only if the two speech signals are equally worded.

We have developed a new approach which makes pattern matching applicable for quasi text-independent tasks. Our method first seeks phonetically matching segments in two speech signals as illustrated in figure 3. For these segments we compute the probability that they were uttered by the same speaker. For both subtasks we use neural networks (see [6]). We have shown that a combination of our pattern matching approach and a standard GMM system is much better than either of the systems alone (cf. [7]).

This project is done in framework and with support of the NCCR IM2.

Recent Publications

- [1] H. Romsdorfer and B. Pfister
Text analysis and language identification for polyglot text-to-speech synthesis
Speech Communication (Elsevier), 49(9): 697–724, September 2007
- [2] H. Romsdorfer and B. Pfister
Character stream parsing of mixed-lingual text
In ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (MultiLing 2006), Stellenbosch (South Africa), April 2006
- [3] R. Beutler
Improving Speech Recognition through Linguistic Knowledge
PhD thesis, No. 17039, Computer Engineering and Networks Laboratory, ETH Zurich, January 2007
- [4] R. Beutler, T. Kaufmann, and B. Pfister
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), November 2005
- [5] T. Kaufmann and B. Pfister
An HPSG parser supporting discontinuous licenser rules
In International Conference on HPSG, July 2007 (to appear)
- [6] M. Gerber, T. Kaufmann, and B. Pfister
Perceptron-based class verification
In Proceedings of NOLISP (ISCA Workshop on non linear speech processing), Paris, May 22-25 2007
- [7] M. Gerber, R. Beutler, and B. Pfister
Quasi text-independent speaker verification based on pattern matching
In Proceedings of Interspeech, pages 1993–1996. ISCA, 2007

(accessible from www.tik.ee.ethz.ch/~spr/publ_spg.html)