SPOKEN LANGUAGE IDENTIFICATION - A STEP TOWARDS MULTILINGUALITY IN SPEECH PROCESSING

Jirí Navrátil
Human Language Technologies Group, IBM Research, T.J. Watson Center P.O. Box 218, Yorktown Heights, NY 10598, email:jiri@watson.ibm.com

1 Introduction

With the trend in globalizing the communication technology and providing services to a wide, multilingual public, the ability of machines to distinguish between languages has become increasingly important. Automatic language identification (ALI) finds its potential applications in multi-language spoken dialog systems (such as information terminals), database- and archive-search, retrieval systems, as well as in the human-human communication (call-routing, automatic translation).

During the past decade intensive research efforts have been made covering different solutions for ALI based on various features contained in the speech signal. This paper deals with the problem of a practicable language identification and discusses its current state, approaches and future directions.

The document is organized as follows: Section 2 describes the language characteristics and their use in current ALI systems. Section 3 presents a multi-component solution based on several information sources followed by experimental results in Section 4. Section 5 deals with the question of the relevance of individual features for distinguishing languages and describes a unique perceptual experiment carried out with human listeners. An outlook and future directions for ALI are given in Section 6.

2 Approaches to Language Identification

Several sources of language-discriminative information have intuitively been addressed in the literature as relevant for the task of language identification: the prosody, the acoustics, and the grammatical and lexical structure.

Due to the high complexity of the lexical and grammatical component recent systems have typically backed off to more simple acoustic-prosodic information or used derived lexical features in order to represent the phonetic structure in a less expensive way. Besides systems with prosodic and acoustic features [4][5] aiming at extracting typical melodic and pronunciation patterns, a very promising and feasible way of acquiring language-specific information is the modeling of statistical dependencies inherent in phonetic chains - the phonotactics. In the statistical sense, phonotactics can be viewed as a subset of grammatical and lexical rules of a language. Since these rules differ among languages the discriminancy is naturally reflected in the phonotactic properties. Several contributions were published dealing with the use of phone n-grams, particularly bigrams, for phonotactic modeling and classification of languages [14],[15], [7]-[11],[2]. Various configurations of

---

1This work has been done in the framework of research activities at the Technical University of Ilmenau, Germany, during 1995-1998.
multiple language-dependent phone-recognizers, run in parallel, were designed to represent
different phone repertoires of the languages and to improve the performance of simple
phonotactic components. In [8] a double-bigram decoding architecture was introduced
which employs a single multilingual decoder with separate sets of language models within
and outside the phone-recognizer, which outperformed the parallel architecture and reduced
the computational expense.

Despite the high efficiency of the phonotactic features it is obvious that only the way of
incorporating multiple sources of knowledge can lead to the robustness necessary for prac-
tical applications. In [14] taking the phone durations into account proved to increase the
identification accuracy. Another combination, namely implicit phonotactic and acoustic
modeling, was done in [15] by using several language-dependent phone-recognizers with
implicit language models where the resulting acoustic likelihoods were taken for the final
decision. In both cases the computational costs were considerable due to the multi-
ple recognition process. An efficient phonotactic-acoustic system that combines extented
phonotactic and acoustic-prosodic features using a neural-network classifier was proposed
in [11] and will be described more in detail in the following section.

Beyond systems based on phonotactics, several studies on ALI as a “byproduct” of
Large Vocabulary Continuous Speech Recognition (LVCSR) were published [3][12]. These
approaches employ full-fledged language-specific LVCSR systems to decode the speech and
use the system with the highest resulting likelihood score for identifying the language.
While reaching the best performance due to the incorporated lexical knowledge, these
approaches incur a considerable amount of computational and training expenses and seem
to be applicable only in systems with few languages and in systems where the LVCSR
results can be reused in further processing (e.g. dialog systems).

3 A Phonotactic-Acoustic System

To make ALI feasible for real applications, which usually require real-time capability, a
compromise between full lexical decoding and a simple phonotactic principle must be found.
As an example of an effective solution, a phonotactic-acoustic system with extended models
is described here.

The system consisting of two principal parts - the phonotactic and the acoustic compo-
nent - is shown in Fig. 1. In the phonotactic component a multilingual phone-recognizer
is used to decode the speech represented by a feature vector sequence. During this process
several language-dependent bigrams (1st-order phone-statistics) are used to weight the
phone transitions in the Viterbi trellis thus producing several ($M$) phone-streams at the
output. Each phone-stream is connected to a separate set of $N$ (outer) language mod-
els capturing the respective phonotactic properties. The core of the language models are
bigrams combined with statistical binary-decision tree models [9]. The latter extend the
modeling context of bigrams by exploiting dependences of up to five subsequent phones and
are based on the minimum prediction entropy criterion. The advantages of this phonotactic
architecture are discussed in detail in [8] [9].

The result of the phonetic decoding, namely the label sequence and the segmentation
information is used by the acoustic component (2) to capture the language-specific phone
pronunciation patterns from the feature sequence using Gaussian mixture models. Since
the acoustic modeling is done on a single decoded phone sequence the computation is
considerably lower than that of the implicit acoustic modeling described in [15]. In addition,
Figure 1: ALI system overview: (1) - Phonotactic block, (2) - Acoustic block.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>6-Lang. 10s</th>
<th>6-Lang. 45s</th>
<th>9-Lang. 10s</th>
<th>9-Lang. 45s</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC-Component</td>
<td>48.9%</td>
<td>48.3%</td>
<td>51.3%</td>
<td>47.8%</td>
</tr>
<tr>
<td>PT-Component</td>
<td>12.8%</td>
<td>3.3%</td>
<td>22.6%</td>
<td>9.4%</td>
</tr>
<tr>
<td>Combined (MLP)</td>
<td>9.8%</td>
<td>0.8%</td>
<td>14.7%</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

Table 1: Error rates on 10/45s utterances in the six- and nine-language-task (NIST’95)

the language-dependent segment duration was included in the acoustic feature vector thus incorporating also prosodic information.

Given a test utterance, both components generate probabilistic likelihoods for each language (phonotactic and acoustic hypotheses). In order to combine the two kinds of hypotheses, the scores are fed into a final classifier for final decision. A multilayer perceptron proved to be superior for this task outperforming linear score combination followed by maximum-likelihood decision.

4 Experimental Results

The described system was tested using a closed set of six and nine languages from the NIST March’95 evaluations involving telephone-quality speech in English, German, Hindi, Japanese, Mandarin Chinese, Spanish, plus three other languages completing the nine language task: French, Tamil, and Vietnamese. Table 1 shows the identification error rates for the two language tasks and two test lengths: 10 sec and 45 sec utterances.

Despite the relatively poor accuracy of the acoustic models used in isolation, combining this information with the phonotactic component consistently improves the overall performance. The unproportional decrease in the phonotactic error rate of the nine language task as compared to six languages is due to the fact that there were no corresponding (inner) bigram models in the Viterbi decoder for the additional three languages. In this case when adding the acoustic information a greater improvement as compared to the six-language task can be observed. Measured on the same evaluation data, the described system out-
performs the multiple-decoder phonotactic-acoustic system [14][13] while keeping the computational expenditure relatively low. This latter advantage has been demonstrated in the development of a real-time on-line language recognizer (BABYLON) [10]. Experimental details can be found in [11][7].

5 Perceptual Experiment

The published results on different ALI systems that involve multiple information sources expose an unequally distributed importance of individual features, in particular the phonotactics, acoustics and prosody, in terms of their performance. Typically the performance decreases in the order of the features as mentioned.

To obtain comparative benchmarks for the significance ranking apparent in ALI, an experiment with human listeners was carried out based on telephone speech signals modified so as to emphasize different information components. A previous perceptual study on language identification was conducted in [6]. Whereas [6] established benchmarks in terms of absolute performances and considered full speech signals only, the experiment presented here aims at measuring the relative importance of lexical, phonotactic-acoustic and prosodic features for the human listener. A differing distribution observed for human as opposed to machine could be an evidence for some discrimination potential that has not been revealed in the current ALI algorithms yet.

The experiment was conducted as a listening quiz in five languages: Chinese, English, French, German, and Japanese and consisted of three test sections. The listeners were asked to identify the language in each test. In the first section (A) the test person could hear speech excerpts (3 and 6 sec long) in each language without any signal modification thus providing the listener with the complete information and serving as a reference. The test signals in the second section (B) consisted of short segment (syllables) which were concatenated in a random order (6 sec). By shuffling syllables, the meaning of the original sentence as well as its prosodic contour were destroyed leaving an equivalent of the phonotactic-acoustic information only. The signals in the third section (C) were filtered by an adaptive inverse LPC filter thus flattening the spectral shape and removing the vocal tract information. This way only the F0 and amplitude components, i.e. the prosody, were audible.

During the collecting period of three months a total number of 78 participants of 12 different nationalities contributed to this study. The overall identification rates for the individual test sections are shown in Table 2.

<table>
<thead>
<tr>
<th>Section</th>
<th>German</th>
<th>English</th>
<th>French</th>
<th>Japanese</th>
<th>Chinese</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>98.7%</td>
<td>100.0%</td>
<td>98.7%</td>
<td>81.7%</td>
<td>88.7%</td>
<td>93.6% (3s)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96.0% (6s)</td>
</tr>
<tr>
<td>B</td>
<td>79.7%</td>
<td>98.7%</td>
<td>79.1%</td>
<td>54.6%</td>
<td>57.7%</td>
<td>73.9% (6s)</td>
</tr>
<tr>
<td>C</td>
<td>32.1%</td>
<td>34.3%</td>
<td>69.4%</td>
<td>45.3%</td>
<td>65.9%</td>
<td>49.4% (6s)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50.0% (15s)</td>
</tr>
</tbody>
</table>

Table 2: Identification rates over all test persons for the sections A (full speech), B (syllables), C (prosody)
The overall ID-rates seem to confirm the relative performance of phonotactic-acoustic and prosodic components in current ALI systems [2] [14][11]. The weakest feature appears to be the prosody - presumably due to its linguistic complexity and the influence of various factors (speaker, emotion, sentence modus, semantics) which also explains the low success achieved with prosody in ALI hitherto [2]. According to the results, the prosody appears to be useful merely for distinguishing between certain broad language groups, e.g. nasal (French), tonal (Chinese), and others, which could be used for language pre-classification.

Despite the relatively good efficiency of the phonotactic-acoustic features (Sec. B), the word-level information in Sec. A seems to be a decisive factor for an acceptable performance. This is an indicator for the necessity of (at least partial) lexical decoding to be included in an ALI system (see discussion below).

Further details on this experiment can be found in [7].

6 Outlook

The problem of automatic language identification gradually gains more interest and application potential. Today's demands for solutions are located especially in information retrieval and audio-archive search engines with access to multilingual documents. Also, applications in the emergency call services [5] (call routing) and radio channel indexing remain relevant. For future dialog-system applications appropriate scenarios of implementing ALI into the dialog still remain to be developed. The questions of how to initiate the dialog in a framework of a multilingual system, how to design possible answers when the confidence in the ALI result is low as well as how to reject unknown languages represent just few examples of the difficulties associated with the application in human-machine communication.

Due to the fact that the computational expense and training requirements connected with the use of LVCSR-based systems (as mentioned in Sec. 2) represent too high a burden for many applications, alternative solutions compromising between full lexical analysis and feasibility can be expected in the future. This trend is becoming obvious in recent publications dealing with keyword-spotting techniques and their use for improving the performance of underlying phonotactic-acoustic systems [1].

Future research attention should be also directed to language rejection, which despite its importance is an underrepresented issue in the literature, as well as to possible links to foreign-accent and dialect identification. Since the boundaries between the definition of a dialect and a language are rather "soft" it can be expected that many algorithms developed for ALI might be useful for accent identification and adaptation in speech recognition.

References


Speech, and Signal Processing (ICASSP), volume 2, pages 1111–4, Munich, Germany, April 1997. IEEE.


