

RESUMO

Identificação automática da língua falada consiste no problema de identificar a língua de uma locução dita por um orador desconhecido, utilizando métodos computacionais. Os sistemas de identificação automática da língua actuais variam muito na sua complexidade. Naturalmente, os sistemas que recorrem a informação linguística de nível superior têm um melhor desempenho. Contudo, essa informação é extremamente difícil de recolher para cada nova língua. Neste trabalho, apresentamos um sistema de identificação da língua falada, com desempenho ao nível dos melhores sistemas publicados, que recorre a pouca informação linguística, sendo por isso fácil de estender a novas línguas. De facto, o sistema apresentado necessita apenas de um reconhecedor de fones de uma única língua (no nosso caso o Português), sendo treinado com sinais de fala de cada uma das outras línguas a identificar.

Estudámos a identificação da língua no contexto das línguas europeias (incluindo pela primeira vez o Português europeu), o que nos permitiu observar o efeito da proximidade linguística entre línguas da família indo-europeia. Os resultados confirmam que as proximidades linguísticas têm um impacto significativo na identificação. Para o corpus SpeechDat-M, com 6 línguas Europeias (Alemão, Espanhol, Francês, Inglês, Italiano e Português) o sistema atinge uma taxa de identificação de cerca de 80%, com locuções de 5 segundos de duração média, o que o coloca ao nível dos melhores sistemas publicados.

ABSTRACT

Automatic spoken language identification is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. Current language identification systems vary in their complexity. The systems that use higher level information have the best performance. Nevertheless, that information is hard to collect for each new language. In this work, we present a state of the art language identification system, which uses very little linguistic information, and so easy to extent to new languages. In fact, the presented system needs only one language specific phone recogniser (in our case the Portuguese), and is trained with speech from each of the other languages.

We studied the problem of language identification in the context of the European languages (including, for the first time, European Portuguese), which allowed us to study the effect of language proximity in Indo-European languages. The results reveal a significant impact on the identification of some languages. With the SpeechDat-M corpus, with 6 European languages (English, French, German, Italian, Portuguese and Spanish) our system has an identification rate of about 80% on 5-second utterances.