Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 781–790

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Talker discrimination across languages ☆

Mirjam Wester *

The Centre for Speech Technology Research, The University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom

Received 23 June 2011; received in revised form 19 January 2012; accepted 23 January 2012
Available online 9 February 2012

Abstract

This study investigated the extent to which listeners are able to discriminate between bilingual talkers in three language pairs – English–German, English–Finnish and English–Mandarin. Native English listeners were presented with two sentences spoken by bilingual talkers and were asked to judge whether they thought the sentences were spoken by the same person. Equal amounts of cross-language and matched-language trials were presented. The results show that native English listeners are able to carry out this task well; achieving percent correct levels at well above chance for all three language pairs. Previous research has shown this for English–German, this research shows listeners also extend this to Finnish and Mandarin, languages that are quite distinct from English from a genetic and phonetic similarity perspective. However, listeners are significantly less accurate on cross-language talker trials (English–foreign) than on matched-language trials (English–English and foreign–foreign). Understanding listeners' behaviour in cross-language talker discrimination using natural speech is the first step in developing principled evaluation techniques for synthesis systems in which the goal is for the synthesised voice to sound like the original speaker, for instance, in speech-to-speech translation systems, voice conversion and reconstruction.

Keywords: Human speech perception; Talker discrimination; Cross-language

1. Introduction

The ability to recognise a person as an individual based on their voice is something most of us probably take for granted. However, if that same individual was speaking a different language – one we maybe did not understand – would we still be able to recognise them as the same person? In most everyday situations this is not a naturally occurring scenario. But, with the advancement of speech technology, scenarios like this are becoming a reality and the question does arise. In the EMIME project,[1] the goal was personalised speech-to-speech translation (S2ST) such that a user's spoken input in one language is used to produce spoken output in another language, while continuing to sound like the user's voice (Wester et al., 2010). This objective raises a number of questions: What is the effect of the modelling techniques used? How to measure whether a person sounds like the same person across language boundaries? How does comparing natural and synthetic speech impact on this?

A survey of the literature in voice conversion – which can be seen as related to speaker adaptive cross-lingual speech synthesis – gives an impression of the types of evaluation that are most commonly used in the field. Abe et al. (1991) used bilingual data (Japanese/English) and measured similarity by calculating mutual information between speaker pairs. Mashimo et al. (2001) also used bilingual data (Japanese/English) and used the objective measure Mel Cepstral Distortion (MCD) to evaluate speaker individuality. In the S2ST project TC-STAR (Sündermann et al., 2006) data from monolingual speakers was used in

☆ Preliminary reports of this work were presented at Interspeech 2010 (Wester, 2010a) and are due to be presented at Interspeech 2011 (Wester and Liang, 2011a).

* Tel.: +44 (0) 131 650 4434; fax: +44 (0) 131 650 6626.
E-mail address: mwester@inf.ed.ac.uk

[1] http://www.emime.org.

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2012.01.006

a unit selection system. Evaluation was carried out using mean opinion scores (MOS) for similarity and quality. The work of Latorre et al. (2006) has a slightly different focus: multilingual synthesis, which is the ability to generate utterances in more than one language, or utterances of mixed language, from a single system. They also use MOS, for intelligibility, similarity and native accent.

A common technique, used in several of these studies, is to compare cross-lingual voice conversion to intra-lingual voice conversion. However, this does not directly measure how similar the speech sounds to that of the original speaker. Using mean opinion scores to evaluate similarity, although a widely-used technique, is not without problems: judging how similar utterances are on a scale from 1 to 5 may be too difficult for listeners, especially if the utterances are in different languages. The results in (Liang et al., 2010) support this. Judgements of speaker similarity are also strongly correlated with the overall quality or naturalness of the synthetic speech: listeners are probably unlikely to rate an utterance as sounding like the target speaker if the quality is poor.

In summary, the methods commonly employed to evaluate talker similarity for voice conversion are no more sophisticated than those already used to evaluate text-to-speech (TTS). Whilst listening tests based on pairwise comparisons or MOS ratings are simple to administer and analyse statistically, they offer no guarantee that what is being evaluated really is talker similarity, independent of other factors such as quality or naturalness.

There are issues associated with comparing synthetic to natural speech, for instance, synthetic speech is less intelligible than natural speech, it requires more cognitive resources, and it is more difficult to comprehend (Winters and Pisoni, 2005). However, there is a more fundamental question which has not been addressed in the voice conversion and speaker adaptation TTS fields which is: to what extent are listeners able to judge talker similarity across language boundaries? This study focuses on this more fundamental question and investigates how well listeners judge talker similarity across language boundaries using stimuli that consist of natural speech.

What are listeners doing when they recognise a talker? To start with there is a distinction that can be made between *what* somebody says and *how* they say it. The *what* is covered by the linguistic properties of speech, that is the message that the speaker is trying to convey, and the *how* is covered by the characteristics of the talker (age, gender, emotional state, health, etc.), i.e. the non-linguistic information or the indexical properties. One of the main questions that has been addressed in previous studies is whether or not there is perceptual integration of these indexical and linguistic properties or if they are independently processed (see e.g., Nygaard, 2005; Winters et al., 2008, for reviews).

Nygaard (2005) gives a comprehensive overview of the relationship between linguistic and non-linguistic information in spoken language processing. She argues, based on the available evidence, that linguistic and non-linguistic information are integrally related components of the same acoustic speech signal and consequently the speech perceptual process.

Neuroscientific evidence supporting the integration of the linguistic and non-linguistic information is given in (Perrachione et al., 2009). Listeners without any familiarity of a particular foreign language appear significantly impaired in achieving native-like accuracy at identifying voices speaking that language, even after substantial training (Perrachione and Wong, 2007). Furthermore, Perrachione and Wong (2007) found that although English subjects improved at a task of Mandarin speaker identification they never came to perform as well as native speakers.

Winters et al. (2008) investigated the extent to which language familiarity affects a listener's perception of the speaker-specific properties of speech by testing listeners' identification and discrimination of bilingual talkers across German and English. They showed that listeners can generalise knowledge of talkers' voices across these two phonologically similar languages. However, it is unknown whether this is also the case for languages that are less closely related. Winters et al. (2008) concluded that listeners apparently process indexical information in a language-dependent fashion when they hear a language that they know; otherwise, they perform indexical tasks by more heavily relying on language-independent information in the signal.

An important factor in speaker identification or discrimination is talker familiarity. Whether or not a listener is familiar with a talker will influence how well they can recognise or identify them, as well as how well they can discriminate between them and other talkers (Kreiman and Papcun, 1991; Van Lancker and Kreiman, 1987). Of course unfamiliar voices can become familiar voices with training. In (Nygaard and Pisoni, 1998) talker-specific learning in speech perception was investigated. They found that listeners' familiarity with talkers facilitated speech intelligibility and that listeners learned talker identity faster from sentences than from single words.

In addition to talker familiarity a listeners' familiarity with the languages under consideration is also of interest. Goggin et al. (1991) investigated talker identification performance in a foreign versus native language. Native English listeners identified bilingual talkers speaking either English or German. Goggin et al. (1991) found that listeners are better at this task when the talkers are using the listeners' native language than when speaking a foreign language. Similar findings have been reported in (Philippon et al., 2007). There it was shown that ear-witnesses are more accurate at recognising voices speaking their native language than an unfamiliar language.

Stockmal et al. (2000) investigated whether listeners are able to separate talker voice from language characteristics, and found that they are able to make same-language/different-language discrimination judgements at better than chance levels. However, in (Stockmal et al., 2004) when asked to focus on voice quality to judge voice similarity in a foreign language, monolingual listeners were not able to ignore language characteristics.