Man vs. Machine

Comparing cross-lingual automatic and human emotion recognition in background noise

Jiska Koemans

Research Master's thesis *Linguistics and Communication Sciences*

Radboud University TU Delft UNIVERSITY OF TWENTE.

Table of contents

Summary

Automatic emotion recognition (AER) from speech has seen major advancements the past decade, but more research is necessary to gain a better understanding of AER's possibilities and pitfalls. AER performance is still impeded by the presence of background noise or in multilingual/cross-lingual circumstances. Especially the combination of these two adverse listening conditions, i.e. cross-lingual emotion recognition in noise, has not yet been taken into consideration in AER studies. In this study, I compare machine performance on speech emotion recognition with human performance on the same task. I take human performance as an upper-bound because humans still outperform machines in emotion recognition, and even in adverse listening conditions human emotion recognition (HER) remains good. Specifically, I investigate the impact of noise and/or an unknown language on AER and compare this to HER in the same adverse conditions. I also investigate which acoustic features play a role in cross-lingual AER in noise, and compare these to their role in cross-lingual HER in noise. Results showed that cross-lingual AER performance was overall lower than cross-lingual HER performance. Cross-lingual AER performance was best for sadness but did not reach chance-level for fear, joy, and anger, and AER performance differed substantially from the cross-lingual HER results. The presence of noise did not have an influence on cross-lingual AER. The analysis of the acoustic parameters showed differences between the parameters linked to recognition of anger and sadness in AER compared to HER, while the acoustic parameters associated with recognition of joy and fear were almost identical in AER and HER. The findings of this study outline the differences that (still) exist between AER and HER, but also observe some similarities. These findings emphasize the importance of comparisons between AER and HER, to be able to better investigate, explain and improve AER, especially in challenging circumstances.

1. Introduction

Emotion perception is an important aspect of everyday communication. Acoustic and non-acoustic aspects of emotion perception help us in the process of correctly identifying an interlocutor's message, and research has shown they both contribute to the process in unique ways (e.g., Castellano, Kessous, & Caridakis, 2008). Especially non-acoustic aspects of human emotion recognition (HER from hereon), such as facial expressions, have received much attention in the literature, but the importance of acoustic properties of emotions for the recognition process has long been established as well (Banse & Scherer, 1996; Scherer, 1986).

In our current society, automatic speech recognition (ASR from hereon) is applied extensively for commercial, health and academic purposes, such as customer services, apps like Siri, or intelligent robots. However, to ensure effective communication, adequate automatic emotion recognition (AER from hereon) in speech – a sub area of affective computing – is of great importance as well, as miscommunications are bound to arise when an ASR system cannot also (correctly) identify a user's emotions (see e.g. ten Bosch, 2003). More recently AER has gained increasing attention in the field of both linguistics and machine learning (Peter & Beale, 2008; ten Bosch, 2003). AER entails the programming of machines to automatically identify emotions from a speech signal, by training them on large sets of emotional speech data through which they 'learn' the specific acoustic characteristics associated with certain emotions. The acoustic feature patterns that are learned through training are then mapped onto the patterns of newly presented test data, to identify which emotion is being conveyed.

Despite the shared goal of HER and AER (i.e., perception of (acoustics of) emotional speech), they differ in terms of their approaches and the difficulties that arise during the recognition process. We do not yet completely understand the underlying processes of AER and HER, or how these are influenced by specific communicative circumstances (e.g. background noise, or different languages). Moreover, human performance seems to be more robust in challenging circumstances than machine performance, but this observation currently cannot be confirmed by the literature because direct comparisons between HER and AER are seemingly lacking. In order to gain a better understanding of both AER and HER, comparing them in the same experimental conditions can help. Such comparisons can for instance provide information

on where obstacles arise in AER and HER and whether these are similar or different; or how AER and HER make use of information in a speech signal, information which in turn can be used for improving AER. Moreover, information on cross-lingual HER can be useful for determining the language- or culture-specific rules that should be implemented in AER as well. The current study will provide a comparison between AER and HER in similar experimental circumstances.

Cross-lingual human acoustic emotion recognition has been studied extensively throughout the years. Many studies have focused on the extent to which human listeners are capable of recognizing emotions in a language they do not speak, specifically when no visual stimuli are present. These studies show that humans are very well capable of this, however, performance does depend on the distance between the investigated languages and the investigated emotions (Banse & Scherer, 1996; Pell, Monetta, Paulmann, & Kotz, 2009; Scharenborg, Kakouros, & Koemans, 2018; Scherer, Banse, & Wallbott, 2001; Scherer, Clark-Polner, & Mortillaro, 2011; Scherer, Wallbott, & Summerfield, 1986; Thompson & Balkwill, 2006; Wallbott & Scherer, 1986).

The presence of background noise also interferes with how well humans can recognize emotions (both within languages and cross-lingually), but to the best of my knowledge only a few studies seem to have focused on the influence of background noise on HER (Parada-Cabaleiro et al., 2017; Scharenborg et al., 2018). Despite the negative impact of background noise that was observed compared to clean, humans were still able to reliably recognize emotions in background noise. This suggests that HER might not suffer severely from the presence of background noise, however, more research on this subject is necessary.

Many studies in the past two decades focused on (improving) automatic recognition of emotional speech, and automatic emotion recognition systems perform quite successfully nowadays, albeit in optimal circumstances (e.g., without any background noise, or when presented with acted emotional speech rather than naturalistic/spontaneous emotional speech) (El Ayadi, Kamel, & Karray, 2011; Schuller, 2018; Schuller, Arsic, Wallhoff, & Rigoll, 2006; Schuller, Lang, & Rigoll, 2002; Schuller, Steidl, & Batliner, 2009; Schuller, Vlasenko, Eyben, Rigoll, & Wendemuth, 2009; Tao & Tan, 2005). However, improvements are necessary in order for AER to

be effectively used 'in the wild', such as for commercial or health applications, like customer services or social robots.

To be able to enhance AER performance, several challenges still need to be overcome, and the current study focuses on two of these: the challenge of cross-lingual AER and the challenge of background noise. First, in the current growing multilingual society machines are likely to be confronted with different languages and therefore it is important that an AER system can deal with multiple languages or cross-lingual emotion recognition. While machines are theoretically capable of cross-lingual emotion perception, accuracies obtained for cross-lingual AER differ depending on the investigated languages, the quality of the speech data in the train and test set, the amount of speakers and/or whether the speakers are actors or not (Feraru, Schuller, & Schuller, 2015; Koolagudi & Rao, 2012). Furthermore, research shows that AER performance is impeded by the presence of noise and despite several attempts to improve AER performance in noisy environments, it is not yet clear what the optimal method would be (e.g., Schuller et al., 2006; You, Chen, Bu, Liu, & Tao, 2006; Zhao, Zhang, & Lei, 2013).

While research has focused on both cross-lingual emotion recognition and the effect of noise on emotion recognition, the combination of these two adverse listening conditions, i.e., cross-lingual emotion recognition in noise, has to the best of my knowledge only been studied once, namely in Scharenborg et al. (2018). Since this study focuses on cross-lingual HER in noise, it seems that the influence of background noise on AER has only been considered in monolingual situations (i.e. training and testing on the same language). Investigating HER and AER in challenging circumstances could help gain a better understanding of their underlying processes. It would allow researchers to gain better perspective of the challenges both AER and HER still face, for instance by outlining which acoustic aspects of a specific emotion are difficult to perceive, to ultimately improve AER and make it better sustainable in challenging circumstances.

Finally, and perhaps most importantly, studies typically do not consider both HER and AER, while such comparisons could provide crucial information regarding the differences that currently exist between the fields. To my knowledge the study by Jeon, Le, Xia, and Liu (2013) is the only existing study that provides a direct comparison between (cross-lingual) AER and HER on the same data set, and they observe better performance for the humans than the machines.

Studies that focus solely on HER often also observe better performance for humans than AER studies do for machines, especially when communicative circumstances become more difficult. These results are however not completely comparable, because such studies almost never consider the exact same experimental conditions and/or emotional speech data. Therefore, one cannot extrapolate HER findings to AER and vice versa. Using the exact same speech data for the investigation of both HER and AER allows researchers to draw a direct comparison between AER and HER, which would help tease apart whether humans and machines make use of the acoustics in a speech signal in similar ways. This would help explain not only a difference in performance, but for instance also why problems occur for AER which typically do not occur for HER. Moreover, because humans currently outperform machines, HER performance can be taken as an upper bound for AER, some sort of goal to achieve – and perhaps ultimately to surpass, if we could develop machines that outperform humans in emotion recognition.

The current study aims to provide a direct comparison between automatic and human emotion recognition on the same data set, in challenging (listening) conditions: cross-lingual recognition of Italian emotions by a Dutch Support-Vector Machine (SVM)/by Dutch listeners, in three listening conditions; one without the presence of background noise (i.e., clean) and two background noise conditions (i.e. SNR +2 dB and SNR -5 dB). As human performance is generally better than AER performance, especially in challenging circumstances, I will take cross-lingual HER performance in noise as an upper-bound (obtained from Scharenborg et al., 2018), against which I will compare AER performance (which is typically referred to as benchmarking in AER literature (e.g., Schuller, Vlasenko, et al., 2009)). I will investigate the role of seven acoustic features in AER, meaning that I will investigate which acoustic features are used by the SVM to identify a specific emotion and how (i.e. is a high pitch associated with more or fewer correct anger responses?). I will compare this to the role of the acoustic features in HER.

I aim to answer the main question *how does cross-lingual AER in noise compare to cross-lingual HER in noise?*, which is divided into the following questions: (1) *how well does a cross-lingual AER model perform in clean speech and in background noise*?, (2) *what is the role of the acoustic features in cross-lingual AER in noise*? and (3) *how does cross-lingual AER compare to cross-lingual HER* (3a) *in its performance* and (3b) *in the role of the acoustic features*?

In the next chapter I will provide an overview of literature on HER and AER. As I will take HER as an upper-bound, I will first discuss HER literature and HER challenges, followed by a discussion of AER literature and the challenges that still exist for AER. In the methods chapter I will then provide a description of the experimental approach of the current study, as well as descriptions of the speech materials that were used and the corpora they were obtained from, and the method of data analysis. In the results chapter will present the cross-lingual AER results and compare these to the cross-lingual HER results from Scharenborg et al. (2018), and in the discussion chapter I will explain the implications of the findings. Finally, I will provide a conclusion of the study.

1.1. Human emotion recognition

1.1.1. Universality of emotions

Human emotion recognition has received attention in the literature for decades (Descartes 1649), with a large focus on facial expressions (Ekman, 1992a, 1992b, 1999; Izard, 1992), but on acoustics as well (Banse & Scherer, 1996; Scherer, 1986)[1]. Darwin (1872/1998) already suggests that emotions contain certain universal aspects. In his book, he focuses on six emotional states: anger, sadness, happiness, fear, surprise and disgust. Ekman (1992a, 1992b, 1999) later proposes the concept of 'basic emotions', with which he claims that certain emotions contain universal aspects that allow them to be recognized across languages and cultures, and importantly, are considered innate and therefore do not need to be learned. The emotions he proposes to be 'basic' are the same six emotions as discussed by Darwin (1872). To this day, these six emotions are most often considered basic and/or universal, despite an ongoing debate as to whether the list should be expanded, with for instance emotional states such as contempt, relief, love and jealousy having also been considered basic (Kowalska & Wróbel, 2017).

[1] I believe it is important to note that many of the studies used participants from so-called 'WEIRD societies': 'Western Educated Industrialized Rich and Democratic societies', which might provide a skewed image of the true universality of emotions (among other things) (e.g. Majid & Levinson, 2010). Indeed, some recent studies that compared Western and non-Western societies have indicated that facial expressions of emotions are not always consistently expressed and perceived across cultures (e.g. Gendron, Roberson, van der Vyver, & Barrett, 2014b; Jack, Garrod, Yu, Caldara, & Schyns, 2012). This is likely also the case for vocally expressed emotions, as studies focusing on this subject come from similar countries. I do however not mean to suggest that emotions do not contain universal characteristics; their universality simply might be less robust than initially thought.

Not only facial expressions have been found to contain universal aspects; research throughout the years has also shown that emotions expressed through speech are recognized across languages and cultures (e.g., Banse & Scherer, 1996; Scherer et al., 2001; Scherer et al., 2011; Scherer et al., 1986; Van Bezooijen, Otto, & Heenan, 1983). Those emotions that have been reliably recognized cross-lingually overlap with the basic emotions as proposed based on facial expressions (Scherer et al., 2011). This is perhaps no coincidence, as it is very likely that this proposed universality hinges on evolutionary purposes (Nesse, 1990). That is, the basic emotions all accommodate some characteristics that could be helpful in certain dangerous or social situations, which in turn might be considered universal as well. For instance, anger and fear are universally important for the fight or flight instinct we and animals intrinsically possess, disgust might be important for the recognition of dangerous or poisonous foods, and correct interpretation of happiness, sadness and surprise contributes to effective communication in social situations (Nesse, 1990). Hence, it is not surprising that universality of these emotions exists in multiple modalities (i.e. facial expressions and vocal emotion expression).

1.1.2. Vocal emotion recognition

With respect to vocal emotion recognition, studies have shown that human listeners are capable of reliably recognizing emotions within and across languages, but results often show that listeners perform best in their own language, sometimes also referred to as an in-group advantage (Elfenbein & Ambady, 2002, 2003). Furthermore, in some cases, languages that are more closely related (e.g., within the same language family versus between language families) are recognized better cross-lingually[2] (Pell et al., 2009; Scherer et al., 2001; Thompson & Balkwill, 2006). So, while vocally expressed emotions indeed seem to exhibit universal characteristics, this 'universality' is limited and emotions are partly language- and/or culture-specific as well.

[2] While the debate on (the difference between) language and culture is a different one entirely, I wish to shortly mention the difference between cross-lingual and cross-cultural emotion recognition. Cross-lingual emotion recognition theoretically refers to emotion recognition across the boundaries of languages, where cross-cultural emotion recognition then refers to emotion recognition across the boundaries of cultures. However, the concepts seem to be closely related and cannot easily be taken apart, as linguistic and cultural elements influence each other and thereby also influence emotion expression and recognition. This study investigates cross-lingual emotion recognition specifically, which is the term I will maintain throughout this thesis.

Language-specific emotional expression obviously consists of the verbal content/vocabulary of a language, but also acoustic information that comes with the language-specific manner of expression (e.g. variation in pronunciation) (Elfenbein & Ambady, 2003; Scherer et al., 2011). Culture-specific emotional expression then has more to do with cultural norms and values that influence how one expresses their emotions. Consider for instance the difference between individualistic and collectivistic cultures: emotion expression in collectivist cultures has been described as more relational, contextualized, and focused on how emotions relate to the 'group', whereas emotion expression in individualist cultures is more subjective, intrapersonal, and focused on how it reflects the individual feelings of the self (Markus & Kitayama, 1991; Mesquita, 2001). Such cultural differences create "emotional languages" that can overlap to an extent, but also differ depending on cultural norms and values, perhaps comparable to linguistic dialects (Elfenbein, Martin, Lévesque, & Hess, 2007). Such aspects that are unique to specific languages and/or cultures contribute in their own ways to how we express and perceive emotions within and across languages.

These language- and culture-specific aspects of emotion expression nonetheless have not prevented experimental studies from observing universal tendencies in cross-lingual emotion recognition. Many studies focusing on (cross-linguistic) vocal emotion recognition have done so by means of a 'forced-choice paradigm': providing participants with possible answers and asking them to classify the stimuli they hear into one of those answers (i.e. "Which of these five emotions do you hear?"). This has yielded evidence strongly in favour of the idea that emotions may be universal (Pell et al., 2009; Sauter, Eisner, Ekman, & Scott, 2010; Scherer et al., 2001; Thompson & Balkwill, 2006). This experimental paradigm may however also limit respondents in their ability to 'describe' what they hear. It has been observed that participants from an isolated cultural group (i.e. the Himba ethnic group from northwestern Namibia) were able to correctly identify emotional vocalizations expressed by English people when provided with predetermined emotion categories, but a replication study wherein these same participants were asked to describe the emotions freely showed that they did not label them according to those English emotion terms (Gendron, Roberson, van der Vyver, & Barrett, 2014a; Sauter et al., 2010). In other words, the Himba speakers did understand the English concepts that they were provided with,

but when given the opportunity to classify the emotions freely, these concepts were not always similar to or as adequate as their own. On the one hand this finding questions the 'universality' of emotions, and on the other hand it shows that researchers should tread carefully when using forced-choice paradigms to investigate cross-lingual emotion recognition. If not enough overlap exists between the categories in the investigated languages, the results may not be experimentally valid or generalizable, and as a results, the observed 'universality' might be less strong than assumed.

In addition to classification of emotions into specific emotion categories (or discrete emotions, i.e. 'anger', 'sadness', 'fear' etc.) emotions can also be classified along dimensions (Laukka, Juslin, & Bresin, 2005; Posner, Russell, & Peterson, 2005; Russell, 1980). In this type of research it is generally argued that classification of emotions in discrete categories is not (or no longer) adequate, and therefore such studies do not focus on the identification of emotions in terms of emotion labels, but rather following acoustic dimensions or continuous scales such as valence (positive vs. negative) and arousal (active vs. passive) (and possibly also potency, pleasantness and (un)predictability, see e.g. Fontaine, Scherer, Roesch, & Ellsworth, 2007; Goudbeek & Scherer, 2008; Goudbeek & Scherer, 2010). For instance, anger is typically classified in terms of high arousal (active) and low valence (negative), happiness in terms of high arousal (active) and high valence (positive), and fear and sadness in terms of low arousal (passive) and low valence (e.g., Goudbeek & Scherer, 2008).

While the classification of emotions on continuous scales may better allow participants in providing their own interpretations, the absence of strict boundaries in this dimensional approach might exactly be the downside of it, for it also makes distinction between emotions more difficult. Inherently very different emotions can be judged similarly on the same scale, e.g., anger and happiness are similar in terms of their degree of arousal. Combining both approaches might provide the most complete view, because this makes it possible to investigate participants own interpretations of the emotions they are provided with, and it also allows researchers to link dimensional classifications to pre-determined emotional labels (Barrett, 1998; Laukka, 2005; Laukka et al., 2005; Scherer, 2003).

The presence of background noise has been found to interfere in many instances with human speech perception (Garcia Lecumberri, Cooke, & Cutler, 2010) and it has been found to influence human within-language and cross-lingual emotion perception as well (Parada-Cabaleiro et al., 2017; Scharenborg et al., 2018). However, the impact of background noise on human recognition of emotions is still largely understudied, despite background noise oftentimes being present in natural communicative circumstances. Parada-Cabaleiro et al. (2017) observed differences in how several different noise types (i.e. white noise, pink noise and brown noise) affected within-language emotion recognition, with pink noise having the worst impact and brown noise the least.

Scharenborg et al. (2018) investigated the influence of background (babble) noise on cross-lingual human emotion recognition. The study shows that humans were able to cross-lingually recognize emotions in the presence of background noise: even for the most severe noise conditions, recognition rates were obtained that were well-above chance-level. However, noise did have a detrimental effect on recognition compared to recognition in the clean condition. Differences were observed between the investigated emotions, but again even lowest recognition rates were well-above chance-level. It seems a comparison between within-language and cross-lingual HER in noise has yet to be made, but such a comparison would be interesting to further investigate to what extent the decrease in performance observed in Scharenborg et al. (2018) was due to the presence of background noise and/or the interaction between the background noise and the language transfer from Dutch to the unknown language Italian.

1.1.3. Acoustics of emotions

In this study, I am especially focusing on the acoustic aspects of emotion recognition, because in cross-lingual vocal emotion recognition, listeners cannot make use of verbal content to determine what they hear. Rather, listeners must use the acoustic information in a speech signal to identify which emotion is being expressed. Vocally expressed emotions exhibit various combinations of acoustic features (i.e. acoustic profiles) that contribute to emotion recognition and distinction between emotions (e.g. high or low intensity, high or low pitch (F0), more or less variability in intensity/pitch) (e.g. Banse & Scherer, 1996; Goudbeek & Scherer, 2008, 2010; Sobin

& Alpert, 1999). For instance, where anger and joy are both often associated with higher intensity, joy tends to contain higher pitch values than anger, which helps discriminate between the two emotions (e.g., Banse & Scherer, 1996; Goudbeek & Scherer, 2008; Thompson & Balkwill, 2006).

Studies that focus on acoustic expression of emotions have suggested that certain emotions contain similarities in their acoustic profiles across languages (e.g. Scherer, 2000; Thompson & Balkwill, 2006). For instance, in various languages anger is vocally expressed with higher mean intensity and/or a greater intensity range, whereas sadness is vocally expressed with a lower mean intensity/intensity range, and while the acoustic patterns of fear and joy are somewhat more variable, a lower F0 range seems to play a role in fear recognition, and joy is often associated with a higher mean F0 (Goudbeek & Scherer, 2008; Scharenborg et al., 2018; Thompson & Balkwill, 2006).

Emotion recognition from speech is influenced by the quality of the speech signals participants in experimental studies are provided with. For instance, a noisy speech signal will be more difficult to recognize due to the masking properties of background noise (Garcia Lecumberri et al., 2010). Many experimental studies make use of acted emotional speech data, because this type of speech is often of good quality, relatively easy to collect and one can control its content, the recording environment, which speakers are obtained in the data set and so on (Koolagudi & Rao, 2012). However, the use of acted emotional speech also poses problems, because the emotions can be exaggerated, or perceived as over-acted, prototypical or insincere (Campbell, 2000; Wilting, Krahmer, & Swerts, 2006). The authenticity and validity of acted emotional speech has therefore been questioned, which has caused increasing effort to produce corpora of more natural emotional speech, for instance through inducing emotions (Scherer, 2013), but studies have also shown that acted and induced emotions might not differ so much in their usefulness for experimental studies on emotion recognition (Laukka, Neiberg, Forsell, Karlsson, & Elenius, 2011; Scherer, 2013). Nonetheless, the validity of using acted emotions in emotion recognition experiments remains subject to debate.

1.2. Automatic emotion recognition

Emotion-specificity of acoustic features is not only important for human listeners, it is also the basis for AER. In AER studies, automatic recognition systems are trained on emotional (human) speech with the goal to automatically identify the emotion that is being expressed. This is usually done by providing an AER model with a set of acoustic features that is extracted from human speech recordings, whereby the model learns the acoustic profiles that are associated with the specific emotions obtained in the training set. When later provided with data it is not trained on, the system then uses the previously learned features to determine the emotions that are being conveyed in the newly presented test data.

AER systems are for this reason dependent on the information they are trained on. This might limit an AER system in its capabilities, because when an AER system is trained on acoustic information, it can solely make use of this type of information. Humans on the other hand always have additional information available, for instance knowledge about linguistic structures (which can help even when they do not speak the language they hear). When we provide AER systems with combinations of feature sets (i.e. not only acoustics, but for instance also lexical features), research shows enhanced AER performance for combined information sources in comparison to only having a single source available (Lee, Narayanan, & Pieraccini, 2002; Truong & Raaijmakers, 2008). AER systems have also been found to perform better in classifying emotions in terms of dimensions (typically valence and arousal) than emotion categories (El Ayadi et al., 2011).

Background noise masks parts of the acoustics of a speech signal and thereby interferes with AER (Garcia Lecumberri et al., 2010; Schuller et al., 2006), even though AER seems to be less prone to background noise than ASR (Schuller, Maier, & Batliner, 2007). It is important to improve AER performance in noisy environments in order for AER to be applicable commercially, because communication generally occurs in the presence of some sort of background noise. Several methods of improving AER (e.g., speech enhancement algorithms, sparse representation classifiers, large acoustic feature sets) have been investigated and all seem to be successful at least to an extent (e.g., Huang, Guoming, Hua, Yongqiang, & Li, 2013; Schuller et al., 2006; Schuller et al., 2007; You et al., 2006; Zhao et al., 2013). However, these studies are not

comparable in terms of experimental approach and the investigated method, and more research is necessary to determine how AER in noise is best improved.

In addition to interference from background noise, overall variability in the speech signal influences AER performance too: when speech is more naturalistic (e.g. uncontrolled spontaneous speech containing multiple speakers), lower recognition rates have been observed compared to experimentally controlled speech and/or acted speech (Koolagudi & Rao, 2012; Schuller et al., 2007; Schuller, Vlasenko, et al., 2009). Many studies on AER have used databases containing acted speech, which does not resemble speech as encountered in natural conversational circumstances (Batliner et al., 2011; Schuller, 2018; Schuller, Steidl, et al., 2009; Wilting et al., 2006). In order to create AER systems that perform well in natural communicative circumstances, where machines are confronted with speech signals that contain much variation, it is important to develop AER systems that are trained on more naturalistic emotional speech.

Developing AER systems that perform well in more natural communicative settings should also entail the development of AER systems that perform well in multilingual and/or cross-lingual communication, because naturally, in the current multilingual society, AER systems will be confronted with different languages. However, not enough databases exist that are suitable for training and testing of cross-lingual AER systems, because most emotional speech databases do not contain multiple languages, or only contain a limited number of languages (which are typically closely related) (Banziger, Mortillaro, & Scherer, 2012; Feraru et al., 2015; Koolagudi & Rao, 2012). Moreover, it is practically impossible to train an AER system on all the languages in the world. Investigating AER in cross-lingual settings is therefore important to determine whether it is possible to train an AER system on one language such that it can reliably perceive emotions in another (untrained) language too.

If machines would be provided with the 'perfect' set of acoustic features and high quality data, their performance would be optimal. If the quality of the data used to train and test an AER system is insufficient, AER performance drops (Tao & Tan, 2005). Using the 'perfect' set of acoustic features and the highest quality training data is desirable, but often not achievable, because creating new databases is very labour-intensive and readily available databases containing emotional speech data are scarce (Batliner et al., 2011; Koolagudi & Rao, 2012;

Schuller, 2018). Moreover, not all (existing) databases are suitable for AER, for instance because the number of speech samples or speakers obtained in a database is too small (Koolagudi & Rao, 2012). Combining multiple databases to create larger sets of speech samples for AER is often not possible. Most existing databases are recorded under different circumstances, in different recordings studios, and for different (experimental) purposes (Batliner et al., 2011; Feraru et al., 2015). This often makes it impossible to combine databases, and it also makes it really difficult to compare between different AER (and HER) studies, because most studies do not use the same speech samples (Batliner et al., 2011). Additionally, a single database often does not contain speech in multiple languages, which makes it difficult to execute cross-lingual AER studies (Feraru et al., 2015). Those require training data and test data in different languages, which are ideally collected under similar circumstances to minimize potential negative effects that are caused by between-database differences that influence the acoustics of a speech signal.

If you were to compare between two or more AER studies (or AER and HER studies), not only would the speech samples used for training and testing of the AER system need to be comparable (preferably recorded under identical circumstances), the studies would also need to have maintained the same experimental approach. Comparing between existing studies is therefore almost never completely experimentally valid, while such comparisons could provide interesting information regarding the state of the art of AER. Moreover, comparing between AER and HER might provide information from a HER-viewpoint that could help improve AER. For this reason, the current study aims to provide such a comparison, which I will discuss in more detail below.

1.3. Comparing HER and AER

An important gap in the research field of emotion perception is the apparent lack of comparisons between HER and AER. The field of HER thus far seems to have been studied more extensively than AER, and humans seem to outperform machines, especially in challenging circumstances. Direct comparisons between AER and HER should be made in order to be able to draw conclusions on how AER and HER performance differs. Since studies on HER in noise show that human performance is affected but not severely impeded in the presence of background noise,

comparing AER to HER in noise might provide important information on where problems occur for AER, but not for HER. Moreover, cross-lingual HER studies show that humans are able to perceive emotions in languages they do not speak. Comparing cross-lingual AER to cross-lingual HER might provide insights into the 'strategies' that humans use in cross-lingual emotion recognition that should also be implemented in AER to improve performance in multilingual/cross-lingual settings. However, comparing results from existing studies often is not ecologically valid due to methodological differences between studies. For this reason, comparisons should be provided between AER and HER performance on the same data set and in the same experimental conditions, such as cross-lingual AER and HER, AER and HER in noise, and/or a combination of these two conditions.

To the best of my knowledge, only one study has directly compared human and automatic cross-lingual emotion recognition, which is the study by Jeon and colleagues (2013). They found similar recognition rates for HER and AER in within-corpus conditions (i.e. within language), but AER performance decreased more in cross-corpora conditions (i.e. between languages) than HER performance. To my knowledge, no studies have considered cross-lingual AER in background noise, while investigating AER in the combination of these adverse listening conditions may provide interesting insight into the underlying processes of AER. Moreover, both cross-lingual AER and AER in noise are likely to occur 'in the wild', so the combination of these conditions is likely to be encountered as well.

1.4. Aim of the current study

To gain a better understanding of the underlying processes of AER and HER, and to provide insights into possible similarities and differences between them, the current study will provide a direct comparison between automatic cross-lingual emotion recognition in noise and human cross-lingual emotion recognition in noise. The HER results were obtained from Scharenborg et al. (2018), which is a study on cross-lingual human emotion recognition in noise, and is based on my own BA thesis (Koemans, 2016). Based on the results from my BA thesis, in Scharenborg et al. (2018) we adapted the experimental stimuli and noise conditions, and added an acoustic feature analysis to determine which acoustic features played a role in recognition of the

investigated emotions. This set-up will be used in the current study as well. A subset of the data set used in Scharenborg et al. (2018) was used here to ensure comparability between the studies.

The main question of the current study was *how does cross-lingual AER in noise compare to cross-lingual HER in noise?* This question is split up in several smaller questions: (1) how well does a cross-lingual AER model perform in clean and noise?, (2) what is the role of the acoustic features in cross-lingual AER in noise? and (3) how does cross-lingual AER compare to cross-lingual HER (3a) in its performance and (3b) in the role of the acoustic features?

The first research question will shed light on the capability of AER to cross-lingually recognize emotions in noise, a combination of adverse listening conditions that has not been investigated before. The second question then focuses on how specific acoustic features play a role in cross-lingual AER in noise, to create a more detailed image of AER performance in adverse listening conditions. The last question focuses on the accuracies observed for both studies, as well as how the acoustic features used for training and testing contribute to the recognition process, to try to shed light on the processes underlying both AER and HER.

2. Method

In this section I will first describe the Scharenborg et al. (2018) study and its results. Then I will describe the speech data used in the current study for training and testing, as well as the corpora they were obtained from, followed by a description of the SVM train and test procedure. The latter will also include a description of the acoustic features used for training and testing, as well as the acoustic feature extraction procedure and the cross-validation procedure performed to determine the best parameter settings for the models. Finally, the data analysis procedure will be described.

2.1. Human emotion recognition study

In Scharenborg et al. (2018), twenty-four native Dutch participants (4 males; mean age=23.0, SD=4.2) were asked to identify five emotion categories (i.e., anger, fear, sadness, joy, and neutral) from an unknown language (in this case Italian) in a no-noise condition (i.e. clean speech), and two babble noise conditions: SNR +2 dB and SNR -5 dB. The babble noise was composed of eight

neutral Italian utterances from eight speakers (4 male, 4 female), which were originally obtained from the CLIPS corpus (available for download: http://www.clips.unina.it/en/corpus.jsp).In addition, eight acoustic features were extracted from the speech: mean F0, F0 range, F0 variability, mean intensity, intensity range, slope of the long-time average spectrum (LTAS), slope of the MFCC, and the Hammarberg Index (HI). All acoustic parameters were previously found to correlate with the recognition of the investigated emotions (see section 2.3.1 and Scharenborg et al. (2018) for more details).

The results showed that anger was recognized significantly better than joy, fear and sadness; moreover, recognition performance deteriorated in more adverse listening conditions compared to the no-noise condition. Significant effects were found for several of the acoustic features. These will be discussed in more detail in the results section in conjunction with the AER results.

2.2. Speech data

The Dutch emotional speech used for training of the SVM was obtained from the KorEmo corpus (previously called DemoKemo Corpus; Goudbeek & Broersma, 2010; see section 2.2.1.). Italian emotional speech was used to test the models, and was obtained from the stimuli used by Scharenborg et al. (2018), a subset selected from the EMOVO corpus (Costantini, Iaderola, Paoloni, & Todisco, 2014; see section 2.2.2.).

2.2.1. KorEmo corpus

The KorEmo corpus is a database constructed for cross-linguistic emotion perception research, and contains both Dutch and Korean emotional speech (Goudbeek & Broersma, 2010). A single nonsense utterance (i.e. [nuto hɔm sɛpikaŋ]) was constructed following three rules: The nonsense utterance only consists of phonemes that occur in both Dutch and Korean and only contains phoneme sequences that adhere to the phonotactic rules of both Dutch and Korean; the phoneme sequences are meaningless in both languages; and they do not contain any embedded real words (Goudbeek & Broersma, 2010). The corpus contains the following eight emotions: anger, sadness, fear, joy, irritation, pride, relief and tenderness. For each part, eight native Dutch professional actors (four female, four male) and eight Korean professional actors

(four female, four male) uttered the nonsense utterance four times per emotion; as such, 512 utterances were recorded, of which 256 were recorded by the native Dutch actors and 256 were recorded by the native Korean actors.

Of the set of 512 utterances, 128 were ultimately obtained in the final corpus (Goudbeek & Broersma, 2010). Selection of the final set of utterances was done through two judgment studies, which were conducted to determine the quality and naturalness of the Dutch and Korean utterances according to native listeners of each language. Participants were asked to classify the recordings into one of the eight emotional categories that the corpus contained and subsequently rated the utterances in terms of naturalness on a scale ranging from 1 (very unnatural) to 4 (very natural) (Goudbeek & Broersma, 2010). Recognition rates were measured in unbiased hit rates (Wagner, 1993 in Goudbeek & Broersma, 2010), with an unbiased hit rate of > 0.1 indicating sufficient recognition. The two utterances of each actor-emotion pair with the highest unbiased hit rate were selected for the final corpus. This resulted a final set of 128 recordings of eight emotions, of which 64 are produced by native Dutch speakers) and 64 are produced by native Korean speakers).

2.2.2. EMOVO corpus

Italian stimuli were obtained from the EMOVO corpus (Costantini et al., 2014), which consists of 588 recordings of Italian emotional utterances portrayed by six actors, in seven emotional categories: anger, sadness, fear, joy, surprise, disgust and neutral. Six native Italian professional actors (three male (M1, M2, M3) and three female (F1, F2, F3)) recorded these emotions in 14 emotionally neutral utterances, with each actor portraying all utterances. Nine of the utterances are semantically neutral (e.g., 'workers get up early') and five are nonsense sentences (with correct grammar, e.g., 'the strong house wants with bread'). The nine regular sentences consist of two questions, three short sentences and four long sentences. The nonsense sentences consist of three short and two long sentences.

The corpus was initially validated superficially, in the sense that the creators wanted to validate the actors' ability to portray emotions, rather than how well each recorded utterance portrayed the intended emotion (Costantini et al., 2014). To that end, twenty-four native speakers of Italian participated in the validation study, which consisted of a listening task focusing

on all speakers, but only 84 nonsense utterances (two utterances per each of the six actors, for each of the seven emotions) out of the total of 588 utterances. Participants were asked to listen to the provided utterance and then choose from two options which emotion they heard. It was concluded that the actors were all able to portray the emotions, because overall recognition rates were above chance-level (Costantini et al., 2014). Additional validation (Giovannella, Conflitti, Santoboni, & Paoloni, 2009) and acoustic analyses (Giovannella, Floris, & Paoloni, 2012) however showed in more detail that both expression and recognition of the emotions strongly vary, depending on actor and emotion (see Giovannella et al. (2009) and Giovannella et al. (2012) for a more detailed overview of the actors' performances and recognizability).

2.2.3. SVM training data

Table 1 displays the number of train and test samples used for training of the SVM, divided per emotion and condition. The subset of Dutch utterances used for training of the Dutch SVM contained four of the eight emotion categories obtained in the KorEmo corpus: anger, sadness, fear and joy. Because these emotions were used in Scharenborg et al. (2018) as well, this would allow me to compare the current AER results with the previously obtained HER results. Originally, I aimed to include neutral speech in the training subset as well, because this emotion category was also used in Scharenborg et al. (2018), but this was not possible due to the fact that the KorEmo corpus did not contain speech data of this category. This should however not influence comparability between the studies because I simply investigate one less emotion category. I chose to train only on the emotions I would also test on, to minimize confusion between emotions. Furthermore, I did not add noise to the training data, only to the test data.

The KorEmo corpus as described in Goudbeek and Broersma (2010) contains only 64 Dutch utterances in total, which means that I would have only 8 utterances per emotion for training of the Dutch SVM (4 emotions, 8 utterances per emotion, 32 utterances in total), which would not be sufficient. Therefore, the creators of the KorEmo corpus provided me with all 512 utterances originally recorded, of which I selected 128 Dutch utterances: all utterances portraying anger, sadness, fear and joy. The training subset thus contained 128 of the 256 Dutch utterances originally recorded for the *KorEmo* corpus. This however also means that 96 of the 128 Dutch utterances were not obtained in the final KorEmo corpus, because they did not meet

the requirements that were set based on the unbiased hit rates observed in the judgment study described in section 2.2.1. (Goudbeek & Broersma, 2010). However, looking at the results from Goudbeek and Broersma (2010), for each of the four emotions investigated in the current study, at least 50% of the actors were recognized well enough, meaning that the results showed unbiased hit rates > 0.1. For sadness, all eight actors were sufficiently recognized; for anger, six actors were sufficiently recognized; for fear, five actors were sufficiently recognized; and for joy, four actors were sufficiently recognized. All Dutch utterances were used nevertheless, because using only the 'best' utterances would result in too small of a training subset, as well as unequal numbers of utterances per emotion.

2.2.4. SVM test data

For testing of the Dutch SVM on Italian I used the same set of utterances that was used in Scharenborg et al. (2018). The human study used only those utterances from the female and male speaker for whom the highest recognition rates were obtained in (Costantini et al., 2014; Giovannella et al., 2009). In the current study, the neutral utterances were not used for testing: only the utterances containing anger, fear, sadness and joy were used. This resulted in a subset of 80 Italian utterances, (i.e., ten utterances per speaker, two speakers per emotion, four emotions, i.e., 20 utterances per emotion). All sentences were tested in all three listening conditions. So, in total, 240 Italian utterances were used for testing: 80 in the no-noise condition, 80 in SNR + 2dB and 80 in SNR -5 dB.

Finally, because the Dutch utterances from the KorEmo Corpus and the Italian utterances from the EMOVO Corpus were recorded under different circumstances, I downsampled the Italian speech data from 48 kHz to a sampling frequency of 44.1 kHz. The mean intensity of the Dutch speech data was increased with 20 dB (variability was preserved). Other acoustic features of the speech data were similar.

Table 1: number of utterances used for training and testing, displayed per emotion and per condition

	Training (clean)	**Test (clean)**	**Test (SNR +2 dB)**	**Test (SNR -5 dB)**
Language	Dutch	Italian	Italian	Italian
Anger	32	20	20	20
Sadness	32	20	20	20
Fear	32	20	20	20
Joy	32	20	20	20
Total	128	80	80	80

2.3. Procedure

Support Vector Machines (SVM) with radial basis function kernel and C-SVC multi-class classification were trained and tested using the e1071 package in R (LibSVM; Chang & Lin, 2011; R Core Team, 2019). The feature vectors of the emotional utterances used for both training and testing were obtained with an acoustic analysis in Praat (Boersma & Weenink, 2019). Table 2 displays all classification experiments that were performed, including their purpose and the conditions they were tested in (NL refers to Dutch, IT refers to Italian).

First I performed an 8-fold cross-validation with the Dutch emotional speech, wherein a Dutch SVM was trained on seven Dutch speakers and tested on the eighth (i.e. monolingual Dutch). This was done to determine the best fitting parameters (i.e., gamma and cost) and the model with the highest accuracy. This also provided information on the recognizability of each speaker obtained in the training set for the Dutch SVM. The Dutch cross-validation procedure is described in section 2.3.2.

I then performed a 4-fold cross-validation with the Italian emotional speech, wherein the SVM was trained on three Italian speakers and tested on the fourth speaker, to gather information about the recognizability of the Italian emotional speech and the speakers obtained in this set. This information would allow me to control whether a potential deterioration from the final Dutch model on Italian test data (i.e. cross-lingual classification) would be due to the language transfer from Dutch to Italian, or due to an intrinsic difficulty of recognizing the

emotions in the Italian data. The Italian cross-validation procedure is described in section 2.3.2. This section also contains the cross-validation results, because these results were used solely in preparation of the final training and testing, rather than for investigation of the research questions, because of which they are not obtained in the results chapter.

After the cross-validation procedures the final training and testing of the cross-lingual SVM was performed. A Dutch SVM was trained on all Dutch emotional speech data described in section 2.2.3. This Dutch SVM was then tested on all Italian emotional speech data described in section 2.2.4., in clean speech and in the two noise conditions. The accuracies obtained from the cross-lingual training and testing were statistically analysed and compared to the results from Scharenborg et al. (2018). The accuracies and the findings from the statistical analysis are described in the results chapter.

Finally, as 'sanity checks', I created a Dutch and an Italian baseline model, and I created a 'combination' model. The accuracies obtained from testing of these models were investigated to further explore the possibility that the cross-lingual SVM suffers from the language transfer.

Table 2: Overview of all classification experiments

Experiment	**Train**	**Test**		
	Model	Clean	SNR +2	SNR -5
Cross validation (monolingual)	SVM_{NL} (seven speakers)	NL (eighth speaker)	X	X
	SVM_{IT} (three speakers)	IT (fourth speaker)	X	X
Baselines (monolingual)	SVM_{NL} (all speakers, half of recordings)	NL (all speakers, remaining recordings)	X	X
	SVM_{IT} (all speakers, half of recordings)	IT (all speakers, remaining recordings	X	X
Combination model (cross-lingual)	SVM_{NL+IT}	IT	IT	IT
Cross-lingual (main)	**SVM_{NL}**	**IT**	**IT**	**IT**

2.3.1. Acoustic feature extraction

For training and testing seven of the acoustic features used in Scharenborg et al. (2018) were extracted from both the Dutch and the Italian speech: mean F0, F0 range, F0 variability, mean intensity, intensity range, the slope of the long-time average spectrum and the Hammarberg Index. All features were previously found to correlate with the emotions investigated in this study (see e.g.,Chatterjee et al., 2015; Luo, Fu, & Galvin III, 2007; Scharenborg et al., 2018; Schmidt, Janse, & Scharenborg, 2016; Sobin & Alpert, 1999), meaning that each of the seven acoustic parameters has been found contribute to recognition of (one of) the investigated emotions. A feature-extraction script was applied that I used for my lab rotation project that focused on the acoustics of (the investigated) emotions. The feature extraction was done with a custom-made acoustic analysis script in *Praat* (Boersma & Weenink, 2019).

The data obtained from the acoustic analysis needed to be scaled in order for it to be suitable for training and testing, meaning that the data is standardized such that it fits within a predetermined scaling range that the SVM can work with. To be able to determine the best scaling range, I trained and tested two SVMs on the same set of Dutch emotional speech data: one SVM was trained with data that was scaled with the built-in scale function of LibSVM (called SVM-scale; scaling between -1 and 1) and one SVM was trained with data that was conversed to z-scores. The SVM that was trained on data that was scaled with SVM-scale in LibSVM yielded the best recognition results, and thus SVM-scale was used throughout the study.

2.3.2. Cross-validation procedures

To determine the best fitting parameters (i.e. gamma and cost) and the model with the highest accuracy, cross-validation was performed. I trained eight native Dutch listener models with LibSVM in R (because the Dutch emotional speech set contains 8 speakers) and then performed a leave-one-speaker-out 8-fold cross-validation to determine the best fitting parameters (i.e. gamma and cost) and the model with the highest accuracy. Each model was trained on data from seven speakers and tested on data from the one remaining speaker. Dutch cross-validation results are reported in Table 3. This table shows the 49 gamma-cost combinations that were tested, divided per model (where model 1 is the model that is tested on speaker 1, and thus trained on speaker 2 through 8; and model 2 is tested on speaker 2 and trained on the remaining

speakers and so on), with the highest accuracy depicted in bold for ease of reading. A gamma-value of 0.125 consistently yielded the best results and is thus the only gamma-value that is reported; a gamma-cost combination of 0.125-16 most often resulted in the best performance and was therefore used in all further training and testing sessions, and all performance rates reported from hereon are based on this combination. The mean performance rate of the Dutch cross-validation was 64.1%, but accuracies strongly varied between the models (see Table 3). The cross-validation results indicate that for the SVM the recognizability of the speakers obtained in the Dutch emotional speech data differs, which is in line with Goudbeek and Broersma (2010).

To investigate the possibility that a decrease in performance of the SVM would be due to the SVM not being able to transfer from the training language Dutch to the test language Italian, I cross-validated the Italian speech data too. This also provided information regarding the differences between (recognizability of) the speakers obtained in the test set, since Giovannella et al. (2009), Giovannella et al. (2012) and Costantini et al. (2014) reported differences between the recognizability of the speakers obtained in the EMOVO Corpus. A 4-fold cross-validation (4 speakers from the EMOVO Corpus were obtained in the Italian speech set) was performed with a gamma/cost-combination of 0.125 gamma and 16 cost, as was determined in the Dutch cross-validation. Four models were trained on Italian speech from three of the speakers and then tested on the data from the one remaining speaker. Italian cross-validation results are presented in Table 4 (model 1 is the model that was tested on speaker 1 and trained on speaker 2, 3 and 4; model 2 was tested on speaker 2 and trained on the remaining speakers, and so on). This cross-validation yielded an overall performance of 35.6%. Performance is less variable than in the Dutch cross-validation, but also considerably lower, though it remains above chance-level of 25%. This indicates that the Italian speech data might inherently be more difficult to recognize for the SVM than the Dutch speech data.

Table 3: Dutch cross-validation accuracies (averaged over the four emotion classes fear, anger, sadness and joy) per model per gamma/cost-combination, including average score per model

Model 1	**Accuracy**	**Model 2**	**Accuracy**	**Model 3**	**Accuracy**	**Model 4**	**Accuracy**
Gamma/cost	**(%)**	Gamma/cost	**(%)**	Gamma/cost	**(%)**	Gamma/cost	**(%)**
0.125 / 1	**75**	0.125 / 1	56.3	0.125 / 1	**43.8**	0.125 / 1	37.5
0.125 / 2	68.8	0.125 / 2	62.5	0.125 / 2	**43.8**	0.125 / 2	31.3
0.125 / 4	68.8	0.125 / 4	62.5	0.125 / 4	37.5	0.125 / 4	43.8
0.125 / 8	62.5	0.125 / 8	62.5	0.125 / 8	**43.8**	0.125 / 8	**50.0**
0.125 / 16	68.8	0.125 / 16	**68.8**	0.125 / 16	37.5	0.125 / 16	43.8
Average	68.8	Average	62.5	Average	41.3	Average	41.3
Model 5	**Accuracy**	**Model 6**	**Accuracy**	**Model 7**	**Accuracy**	**Model 8**	**Accuracy**
Gamma/cost	**(%)**	Gamma/cost	**(%)**	Gamma/cost	**(%)**	Gamma/cost	**(%)**
0.125 / 1	56.3	0.125 / 1	56.3	0.125 / 1	**62.5**	0.125 / 1	81.3
0.125 / 2	62.5	0.125 / 2	56.3	0.125 / 2	56.3	0.125 / 2	81.3
0.125 / 4	**68.8**	0.125 / 4	62.5	0.125 / 4	50.0	0.125 / 4	81.3
0.125 / 8	62.5	0.125 / 8	68.3	0.125 / 8	56.3	0.125 / 8	81.3
0.125 / 16	**68.8**	0.125 / 16	**75.0**	0.125 / 16	56.3	0.125 / 16	**93.8**
Average	63.8	Average	63.8	Average	56.3	Average	83.8

Table 4: Italian cross-validation accuracies (averaged over the four emotion classes fear, anger, sadness and joy) per model and averaged over all models

Model	**Accuracy (%)**
1	22.5
2	32.5
3	37.5
4	50.0
Average	35.6

2.3.3. Baseline models and 'combination-model'

Results from the cross-validation suggested that the Italian speech data was more difficult to recognize for the SVM than the Dutch speech data. To investigate this in more detail, I created two new models: one that was trained on part of the Dutch emotional utterances, but now with all available speakers included, and then tested on the remaining Dutch emotional utterances; and an Italian model that was tested on part of the Italian emotional utterances with all speakers

included and then tested on the remaining Italian emotional utterances. These models functioned as a Dutch and an Italian baseline model in 'optimal' monolingual circumstances, wherein one would expect performance to be highest because there is no language transfer needed.

To investigate if the availability of a little bit of training material in the right language increases the cross-lingual SVM's performance – which would strengthen the idea that cross-lingual AER is indeed affected by the language transfer – I also trained an SVM on a combination of Dutch and Italian utterances. This model was then tested on Italian utterances from speakers that were not used for training of the SVM, in clean and in noise at SNR +2 dB and SNR -5 dB.

T-tests were performed to determine whether the results from the baseline models differed from each other, and to determine whether the results from the 'combination-model' differed from the accuracy results obtained for the final cross-lingual SVM testing. These findings will be reported in the results chapter in section 3.1.3. and 3.1.3.

2.4. Data analysis

The findings obtained from the training and testing of the final cross-lingual SVM (i.e. the model trained on all Dutch emotional speech and tested on all Italian emotional speech in clean and noise) were statistically assessed using general linear mixed effects models in R (Baayen, Davidson, & Bates, 2008). This specific method was chosen to ensure comparability with the previous study (Scharenborg et al., 2018), wherein the same method was used. Three separate analyses were performed to investigate 1) whether cross-lingual automatic recognition differed across the investigated emotions and noise conditions; 2) the role of the acoustic features in cross-lingual automatic recognition of Italian emotions in noise; and 3) to take a closer look at the role of the acoustic features in cross-lingual automatic recognition of each of the investigated emotions. For each of the three analyses I will also provide a comparison with the results from Scharenborg et al. (2018).

A backwards stepwise regression method was applied, meaning that all possible interactions and effects were included in the first model and the model was then stripped until the model with the best fit remained. Stripping of the model entails the removal of the least

significant effect from the model (starting with the interactions), to check whether removing this effect improves the model fit. If so, this indicates the effect does not explain the variance in the model and the effect is left out of the analysis. If the model fit does not improve by removing an effect, the effect remains in the model and the next least significant effect is removed. This procedure is repeated until the model fit improves no more by the removal of non-significant effects and the model with the best fit is established.

In the first analysis I investigated whether cross-lingual AER performance was significantly different for the investigated emotions and/or noise conditions. This analysis consisted of a model with correctness (1 = correct, 0 = incorrect) as the dependent variable, and emotion (fear on the intercept), listening condition (clean on the intercept) and gender of the speaker (female on the intercept) as fixed factors. Stimulus (the specific utterances used for testing) and speaker (the actors F1, F2, M1, M2 that portrayed the Italian emotions) were added as random factors. Please note that the gender of the speakers in the test set was not specifically interesting for the question of how automatic recognition is influenced by the presence of noise and/or an unknown language. However, in the HER study, gender was added as a factor to the statistical analysis and several effects of gender were observed; therefore, to allow for comparison between the current AER results and the HER results from Scharenborg et al. (2018), I included gender as a factor in the current analyses as well.

In the second analysis I investigated whether the seven acoustic parameters used for training and testing played a significant role in the cross-lingual automatic recognition of Italian emotions in noise. In this analysis, in addition to the factors already included in the previous analysis, the seven acoustic parameters were added as fixed factors: mean F0, F0 variability, F0 range, mean intensity, intensity range, slope of the long-time average spectrum (slope of the LTAS) and the Hammarberg Index. The automatic results were first analysed as a whole, and then in the third analysis they were analysed per emotion.

The third analysis thus consisted of a set of four separate emotion analyses, wherein the role of the acoustic parameters in cross-lingual automatic recognition of a particular emotion was investigated in more detail. Each emotion analysis contained the same factors as the second

analysis (but now without the factor emotion). The four emotion analyses will be discussed separately, in conjunction with the HER per-emotion results.

Initially, I added the factor human/machine as a fixed factor as well, to determine the differences between human and automatic emotion perception on the current data set. However, the HER data set contained 192 utterances per emotion per listening condition, while the AER data set contained only 20 utterances per emotion per listening condition, due to 24 participants being tested versus only one SVM being tested. As a result, the data sets could not be compared in a statistical analysis, and therefore I compared by hand the results from the per-emotion analysis in Scharenborg et al. (2018) with the results from the per-emotion analysis in the current study.

3. Results

In order to answer the question whether cross-lingual AER and HER perform similarly in noise, I will first report the cross-lingual AER accuracies in clean, noise +2 dB and noise -5 dB, which I will then compare to the cross-lingual HER accuracies from Scharenborg et al. (2018). In this section, I will also report the comparison of the baseline models and the combination model described in section 2.3.3.

Subsequently, following Scharenborg et al. (2018), I carried out three sets of statistical analyses on the cross-lingual AER accuracies. The first analysis was carried out to investigate the question whether the cross-lingual AER accuracies significantly differed between the investigated emotions and noise conditions. In the second statistical analysis, I added the acoustic parameters as fixed factors, to determine if and how they contributed to cross-lingual AER performance in noise. The third and final set of analyses consisted of four separate per-emotion analyses, which investigated in more detail the influence of noise, as well as how the acoustic features contributed to cross-lingual automatic recognition of each of the investigated emotions (i.e. anger, sadness, fear and joy). Importantly, I will provide a comparison of the results of each of the analyses with the HER results from Scharenborg et al. (2018).

3.1. AER accuracies

3.1.1. Results of the monolingual baseline models

Table 5 displays the accuracies of the Dutch and Italian baseline models where all speakers were included in the train and test set, for each emotion. The Dutch and Italian accuracies did not significantly differ from each other (t-test; t(24)=.8, p=.5), which indicates that the Italian speech data are not inherently more difficult to recognize than the Dutch speech data.

Table 5: Accuracies (%) of the baseline models per emotion

Emotion	Dutch	Italian
Fear	96.9	95
Anger	100	95
Sadness	100	95
Joy	90.6	80

3.1.2. Cross-lingual AER in noise: accuracies

The overall emotion recognition accuracy of the Dutch SVM when tested on the Italian speech was 28.3% correct, which is just above the chance-level of 25%. Table 6 displays the cross-lingual accuracies per emotion and per listening condition and averaged over all emotions and listening conditions. The accuracies displayed in Table 6 are comparable for the three listening conditions, however, recognition rates drop below chance-level for three of the four investigated emotions. In addition to the accuracies shown in Table 6, Figure 1 shows the recognition patterns for cross-lingual AER and HER, for each emotion in each listening condition (AER results are depicted in blue, HER results are obtained from Scharenborg et al. (2018) and depicted in orange). The presence of background noise resulted in a drop in AER performance for fear in SNR -5 dB and for anger in SNR +2 dB, but the anger accuracy then increases again in SNR -5 dB compared to SNR +2 dB. The accuracy for sadness is actually higher when background noise is present than when no noise is present, and the accuracy for joy increases in SNR +2 dB compared to clean and SNR -5 dB.

The low accuracies that are observed in cross-lingual AER even when no background noise is present suggest that the SVM was not able to transfer from the Dutch emotions it was trained on to Italian emotions. The observation that the accuracies are largely unaffected by the presence of background noise implies a floor effect: because recognition was already too low, the accuracies did not drop any further under the influence of background noise.

I investigated the high accuracies for sadness and the low accuracies for the other emotions by creating a confusion matrix, to find out in more detail how the accuracies are composed.

Table 7 shows the confusion matrix from the Dutch SVM tested on the Italian data. Each percentage indicates how often a specific emotion was predicted for each emotion category. Sadness was chosen most often, regardless of which emotion an utterance contained. This seems to suggest that the Dutch SVM is not able to recognize the Italian emotions well, and rather is 'backing off' to sadness irrespective of the emotion in the Italian utterance. The high accuracies for sadness should therefore be interpreted cautiously.

Table 6: cross-lingual accuracies for each emotion and listening condition and averaged over all listening conditions and emotions

Accuracies (% correct)					
Condition	*Average*	*Fear*	*Anger*	*Sadness*	*Joy*
Average	X	16.7	5	85	6.7
Clean	28.8	25	5	75	5
SNR +2dB	31.3	25	3	90	10
SNR -5 dB	25	0	10	90	5

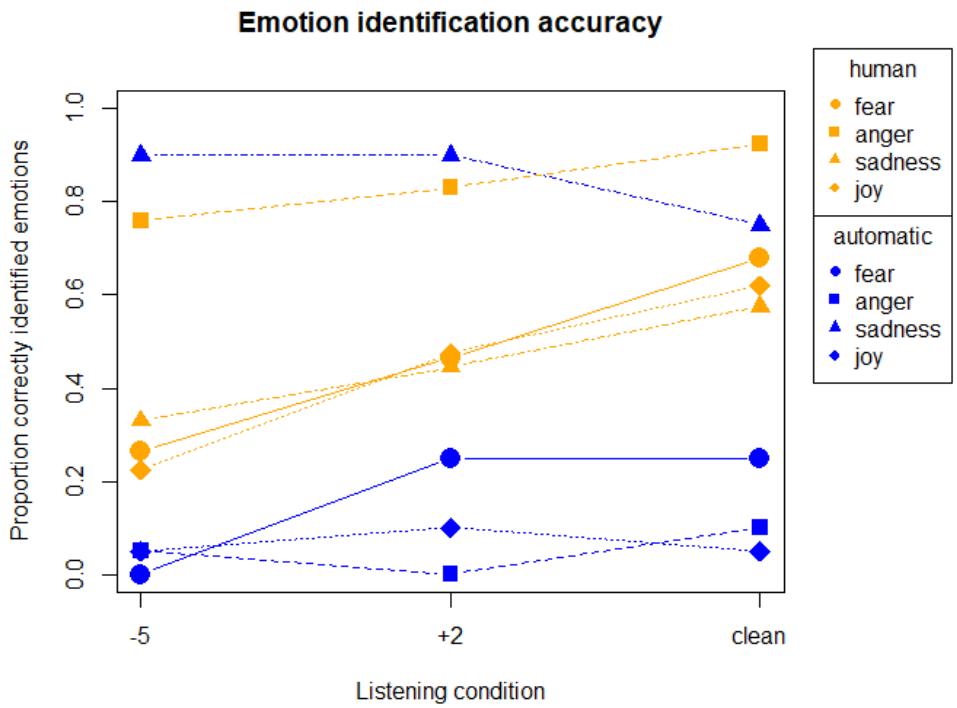

Figure 1: Recognition patterns of AER vs. HER per emotion per listening condition

Table 7: Confusion matrix of percentage (%) predicted emotion per actual emotion category for cross-lingual classification.

Predicted emotion category	**Actual emotion category**			
	Fear	Anger	Sadness	Joy
Fear	16.7	15	3.3	8.3
Anger	13.3	5	10	6.7
Sadness	66.7	73.3	85	78.3
Joy	3.3	6.7	1.7	6.7

3.1.3. Results of the cross-lingual 'combination model'

Accuracies of the 'combination model' (i.e. the model trained on a combination of Dutch and Italian and tested on Italian) are presented in Table 8. These accuracies were compared with the accuracies from the final cross-lingual model in clean and the two noise conditions. When no noise is present, AER benefits from the addition of Italian emotional speech to the training set compared to the final cross-lingual model that was trained on Dutch emotional speech only (t-test; $t(79)$= -3.5, p=<.001). In the presence of noise this advantage was however eliminated (+2

dB: t-test; *t*(79)= .6, p=.5; -5 dB: t-test; *t*(79)= .5, p=.7). These findings show that when the SVM is trained on a little bit of the test language, AER performance increases, which suggests that in this situation the SVM was better able to make the language transfer than in a purely cross-lingual situation (i.e. trained on only Dutch and tested on Italian). However, in the presence of noise, the Dutch-Italian combined SVM does not perform better compared to a purely cross-lingual situation. This suggests that an effect of noise might be present, however, this was not statistically assessed in the current study.

Table 8: Accuracies (%) of the combination model per condition, in comparison with the cross-lingual accuracies

Condition	**Accuracy (combination)**	**Accuracy (cross-lingual)**
Clean	50	28.8
SNR +2 dB	26.3	31.3
SNR -5 dB	25	25

3.1.4. AER vs. HER: comparing the accuracies

Comparing the AER results with those of the human subjects in Scharenborg et al. (2018) (see figure 1) shows that different recognition patterns can be observed for the emotions in AER and HER. Overall, HER accuracies are higher than AER accuracies, and in AER sadness is recognized best, while in HER anger is recognized best. The three remaining emotions in AER (i.e. fear, anger and joy) show lower accuracies than the three remaining emotions in HER (i.e. fear, sadness and joy). Moreover, compared to HER, AER recognition patterns in the noise conditions show much more variation. The HER accuracies all show a linear decline from clean to the noise conditions, independent of the investigated emotion. In AER on the other hand, sadness accuracies in the noise conditions are higher than in clean, and for fear, anger and joy small differences are observed between clean and SNR +2 dB; for these three emotions a decline is observed only in SNR -5 dB.

3.2. The effect of noise and emotion on cross-lingual AER

The first statistical analysis was performed to assess overall cross-lingual AER performance in noise. This analysis did not show any significant effects for noise and gender cross-lingual AER performance, only for emotion category: utterances portraying sadness were recognized significantly better than the other emotions (β=3.89, SE=0.84, p<.001).

3.2.1. AER vs. HER: comparing the effect of noise and emotion

The results from the HER study showed two significant effects: in HER anger was recognized significantly better than all other emotions, and in addition, utterances presented in both noise conditions yielded significantly fewer correct answers from humans than utterances presented in clean. In AER the only significant effect was that of sadness being recognized better than all other emotions.

3.3. The role of the acoustic features in cross-lingual AER

In this analysis the acoustic features were added as fixed factors to the analysis described above, to determine if and how they contribute to cross-lingual AER in noise. Table 9 displays the estimates of all fixed effects in the best-fitting model (significant p-values are displayed in bold for ease of reading). It is important to note that in this analysis, the direction of the main effects cannot be reliably interpreted, because each main effect is also part of an interaction effect. Moreover, the direction of the main effect of sadness that is suggested by the GLM is not in line with the observed accuracies, which indicates that interpretation of the main effects is not as straightforward as simply saying one emotion is recognized better than the other. For this reason I will simply describe which main effects are observed, but refrain from trying to explain how they specifically affect cross-lingual AER to avoid possible mistakes in the interpretation.

All parameters except F0 variability predicted overall automatic emotion recognition, meaning that only F0 variability was nit important for cross-lingual automatic recognition of the investigated emotions. Furthermore, the results indicate that joy and sadness are recognized significantly different than fear, while no effect was observed for anger. Utterances from male

speakers were recognized differently than those from female speakers, and in both SNR conditions, AER was different than in clean speech.

I further observed several interaction effects (see Table 9). A higher F0 mean was associated with more correct answers for joy from the SVM (joy × F0 mean), as was observed for more variability in the F0 (joy × F0 variability) and a greater range of intensity (joy × intensity range). In addition, more variability in the F0 and a greater intensity range were associated with a decrease in correct sadness answers from the SVM (sadness × F0 variability; sadness × intensity range). A higher mean intensity and a higher slope of the LTAS were associated with more correct sadness answers (sadness × mean intensity; sadness × slope of the LTAS). Furthermore, for emotion portrayals from the male speakers compared to female speakers, a higher mean F0 was associated with more correct answers (gender: male × mean F0), while a higher F0 range and mean intensity were associated with fewer correct answers (gender: male × F0 range; gender: male × mean intensity). In background noise at SNR +2 dB, a higher mean F0 and more variability in the F0 were associated with fewer correct responses compared to clean (SNR +2 dB × mean F0; SNR +2 dB × F0 variability), whereas a greater F0 range and a higher mean intensity yielded more correct responses compared to clean (SNR +2 dB × F0 range; SNR +2 dB × mean intensity). At a noise-level of SNR -5 dB, more variability in the F0 was associated with more correct answers compared to clean (SNR -5 dB × F0 variability).

Table 9: fixed effect estimates for the best-fitting model of the overall accuracy analysis including the acoustic features

Fixed effect	**β**	***SE***	***p***
Main effects			
Intercept	171.230	2.467	**.014**
Gender: male	-99.755	43.665	**.022**
SNR +2 dB	11.660	5.527	**.035**
SNR -5 dB	-8.452	4.068	**.038**
Anger	-412.885	1671.282	.805
Sadness	-309.112	121.657	**.011**
Joy	-213.31	79.9	**.008**
Mean F0	-113.491	47.605	**.017**

F0 range	-36.244	13.698	**.008**
F0 variability	25.961	14.463	.073
Mean intensity	-48.855	17.715	**.006**
Intensity range	-55.256	24.052	**.021**
Slope of the LTAS	-26.868	10.824	**.013**
Hammarberg Index	-15.090	6.470	**.020**
Interaction effects			
Anger × mean F0	-41.907	1694.963	.980
Anger × F0 variability	95.167	1201.837	.937
Anger × mean intensity	109.829	750.579	.884
Anger × intensity range	123.462	717.542	.864
Anger × Slope of the LTAS	-172.386	1373.382	.900
Sadness × mean F0	-111.274	82.109	.175
Sadness × F0 variability	-318.441	126.120	**.012**
Sadness × mean intensity	94.120	33.316	**.005**
Sadness × intensity range	-70.079	28.228	**.013**
Sadness × slope of the LTAS	64.133	22.824	**.005**
Joy × mean F0	129.24	55.153	**.020**
Joy × F0 variability	29.753	12.648	**.019**
Joy × mean intensity	41.039	26.516	.122
Joy × intensity range	69.909	27.191	**.010**
Joy × Slope of the LTAS	19.932	13.297	.155
Gender: male × mean F0	509.209	192.154	**.008**
Gender: male × F0 range	-181.064	67.386	**.007**
Gender: male × mean intensity	-58.684	21.537	**.006**
Gender: male × slope of the LTAS	16.479	10.056	.101
SNR +2 dB × mean F0	-29.535	9.719	**.002**
SNR +2 dB × F0 range	53.289	19.854	**.007**
SNR +2 dB × F0 variability	-69.734	25.539	**.006**
SNR +2 dB × mean intensity	32.376	11.440	**.005**
SNR +2 dB × intensity range	-15.481	9.374	.099
SNR +2 dB × Hammarberg Index	1.193	4.482	.790

SNR -5 dB × mean F0	-5.308	3.349	.113
SNR -5 dB × F0 range	-20.429	10.999	.063
SNR -5 dB × F0 variability	25.760	12.282	**.036**
SNR -5 dB × mean intensity	-5.372	3.349	.109
SNR -5 dB × intensity range	3.996	3.856	.300
SNR -5 dB × Hammarberg Index	4.653	2.854	.103

3.3.1. AER vs. HER: the role of the acoustic features in cross-lingual emotion recognition

Table 10 displays all significant effects that were observed in the AER and the HER analyses, including the direction[3] of the effects (indicated with arrows: ↑ means the effect was positive, ↓ means the effect was negative); absent effects are indicated with a hyphen (-), and when an effect was observed for both AER and HER the cells are marked (light green indicates the effects are in the same direction, e.g., both positive; light yellow indicates the effects are in different directions). Comparing the AER and HER results shows that all acoustic features except F0 variability predicted cross-lingual AER, while for only intensity range and none of the other features influenced cross-lingual HER. Altogether, six of the seven investigated acoustic features influenced overall cross-lingual AER, while only one of the features influenced cross-lingual HER.

Furthermore, both AER and HER showed interactions between joy and F0 variability, and between sadness and the slope of the LTAS. The interaction effects are positive in both AER and HER: as F0 variability increases, more correct joy responses are given, and as the slope of the LTAS increases, more correct sadness responses are given. However, the interactions between joy and mean F0, joy and intensity range, sadness and F0 variability, sadness and intensity range and sadness and mean intensity were observed in AER only. Interactions between joy and the Hammarberg Index and between sadness and mean F0 were unique to the HER results.

The AER analysis further yielded interaction effects between emotions uttered by a male speaker and mean F0, F0 range and mean intensity, while the HER analysis did not yield significant interaction effects between gender and any of the acoustic features.

[3] While the direction of the main effects may not be fully interpretable, I still added them to Table 10 in order to provide a thorough comparison. However, as described before, they should be interpreted cautiously.

Finally, the interaction between mean F0 and SNR +2 dB was observed in both AER and HER. The interaction effect is negative in both AER and HER: as mean F0 increases, fewer correct responses are given in background noise at SNR +2 dB. The AER analysis further showed that F0 range and mean intensity interacted with SNR +2 dB, and that F0 variability interacted with both SNR +2 dB and SNR -5 dB. In the HER analysis, additional interaction effects were found between intensity range and SNR +2 dB, and between mean F0 and SNR -5 dB.

In sum, a comparison between the AER analysis containing the acoustic features and the HER analysis containing the acoustic features shows only a few similarities regarding the role of the acoustic features in cross-lingual AER and HER. Only intensity range is associated with both AER and HER. Joy is associated with F0 variability in both analyses, sadness is associated with the slope of the LTAS in both analyses, and mean F0 is associated with SNR +2 dB in both analyses. The direction of each of these interaction effects is also the same in AER and HER. This leaves seventeen additional main and interaction effects that were observed in the AER analysis only, and three additional interaction effects that were observed in the HER analysis only. So, while some overlap exists between the acoustic features found to play a role in AER and HER, they differ for the most part. This may explain the accuracy differences that were observed between AER and HER, which will be discussed in more detail in the discussion section.

Table 10: Overview of all significant effects observed in the overall AER and HER analyses concerning the role of the acoustic features, noise and gender. It is indicated whether an effect was positive (↑) or negative (↓) or absent; marked cells indicate overlap

Main effect	**AER**	**HER**
Mean F0	↑	-
F0 range	↑	-
Mean intensity	↑	-
Intensity range	↑	↓
Slope of the LTAS	↑	-
Hammarberg Index	↑	-
Interaction effect	**AER**	**HER**
Joy × mean F0	↑	-
Joy × F0 variability	↑	↑

Joy × intensity range	↑	-
Joy × Hammarberg Index	-	↓
Sadness × mean F0	-	↓
Sadness × F0 variability	↓	-
Sadness × mean intensity	↑	-
Sadness × intensity range	↓	-
Sadness × slope of the LTAS	↑	↑
Gender: male × mean F0	↑	-
Gender: male × F0 range	↓	-
Gender: male × mean intensity	↓	-
SNR +2 dB × mean F0	↓	↓
SNR +2 dB × F0 range	↑	-
SNR +2 dB × F0 variability	↓	-
SNR +2 dB × mean intensity	↑	-
SNR +2 dB × intensity range	-	↑
SNR -5 dB × F0 variability	↑	-

3.4. The role of acoustic features in cross-lingual AER in noise for each emotion

In order to compare the AER results with the HER results for each emotion separately, separate analyses were carried out on the role of acoustic features, noise and gender in cross-lingual AER in noise (again with generalized linear mixed effect models). Below, I will discuss per emotion the results of each analysis, consisting of the effects observed for the acoustic features, noise, gender and interactions between these factors. An overview of all significant main and interaction effects that were found in both the AER and HER emotion analyses is provided in Table 15 at the end of the results section (similar to

Table 10 in the previous section; arrows indicate the direction of an effect, hyphens indicate the absence of an effect and marked cells indicate that an effect was observed in both AER and HER). Please note that like in the previous analysis, not all observed main effects can be explained due to them being part of interaction effects. For these cases I will solely describe the effect, not its direction.

3.4.1. Fear

Table 11 displays the estimates of the fixed effects in the best-fitting model for the analysis of the automatic recognition of fear (significant p-values are displayed in bold for ease of reading). A main effect was observed for F0 range, which predicted cross-lingual automatic recognition of fear, with a higher F0 range being associated with fewer correct responses. F0 variability influenced AER too in interaction with gender (gender: male × F0 variability).

These results are largely in agreement with the results of the emotion analysis of human recognition of fear. Both F0 range and F0 variability were found to modulate human cross-lingual recognition of fear (Scharenborg et al., 2018), where F0 variability also interacts with gender (gender: male × F0 variability). The HER analysis showed that mean F0 additionally played a role in cross-lingual HER, in general and in interaction with gender (gender: male × mean F0; HER). Mean F0 was not associated with automatic recognition of fear.

Fear portrayals from male speakers were recognized differently by the SVM than those from female speakers. In the presence of noise at SNR +2 dB and at SNR -5 dB, automatic recognition of fear significantly differed from fear recognition in clean speech (see Table 11). Furthermore, for SNR -5 dB only, an interaction with gender was observed, such that utterances from male speakers were recognized less well than utterances from female speakers at a noise level of SNR -5 dB only (SNR -5 dB × gender: male).

The HER results for the effect of noise on fear recognition are in agreement with the results obtained from the statistical analysis for AER: human recognition of fear was different in both noise conditions compared to clean speech, and even more so for male fear portrayals in SNR -5 dB (gender: male × SNR -5 dB; HER). This pattern was also observed in the statistical analysis of automatic fear recognition. Finally, emotions from male speakers were recognized differently than those of female speakers in both AER and HER.

Altogether, F0 variability and F0 range play a similar role in recognition of fear in AER and HER. Additionally, mean F0 plays a role in HER, but not in AER. Background noise has an impact on both AER and HER, and more so for fear portrayals from male speakers than female speakers. Fear portrayals from male speakers are differently than those from female speakers in both AER and HER as well.

Table 11: fixed effect estimates for best-fitting model of the accuracy analysis of fear (AER)

Fixed effect	**β**	***SE***	***p***
Intercept	.820	.239	**< .001**
Gender: male	2.238	.835	**.007**
SNR +2 dB	-.612	.278	**.028**
SNR -5 dB	-1.359	.288	**< .001**
F0 range	-1.363	.511	**.008**
F0 variability	1.693	.704	**.016**
Hammarberg Index	.339	.174	.051
F0 variability × gender:male	3.568	1.212	**.003**
SNR +2 dB × gender: male	-.442	.398	.267
SNR -5 dB × gender:male	-.975	.441	**.027**

3.4.2. Anger

Table 12 displays the estimates of the fixed effects in the best-fitting model for the analysis of the automatic recognition of anger (significant p-values are displayed in bold for ease of reading). Only one acoustic feature was associated with cross-lingual AER: increasing intensity range predicted fewer correct anger responses from the SVM (see Table 12).

The intensity range feature was also important for HER in Scharenborg et al. (2018), where an increased intensity range predicted a decrease in correct anger responses from humans, which was observed for AER too. However, six additional acoustic features were associated with anger recognition in HER compared to AER. An increased mean intensity predicted fewer correct anger responses from humans, while an increased Hammarberg Index predicted more correct anger responses from humans. Furthermore, in HER, an increased mean F0, increased F0 range, increased slope of the LTAS and increased Hammarberg Index were all associated with fewer correct anger responses for fear portrayals from male speakers (mean F0 × gender: male; F0 range × gender: male, slope of the LTAS × gender: male; Hammarberg Index × gender: male; HER). An increased mean intensity, intensity range and F0 variability were associated with more correct responses for fear portrayals from male speakers in HER (mean intensity × gender: male; intensity range × gender: male; F0 variability × gender: male; HER).

Similar to HER, increasingly bad SNRs led to increasingly fewer correct responses for AER (see Table 12). No effects of gender were observed for AER, while for HER, emotions from female speakers were recognized better than those from male speakers.

In sum, only one acoustic feature plays a role AER compared to seven acoustic features playing a role in HER (see Table 15). Intensity range was found to be important in the recognition of anger by both the SVM and the humans. At the same time, HER is modulated by six additional acoustic features that play no role in AER. Noise was found to negatively affect both AER and HER, but only for HER an effect of gender is observed.

Table 12: fixed effect estimates for best-fitting model of the accuracy analysis of anger (AER)

Fixed effect	**β**	***SE***	***p***
Intercept	1.974	.214	**< .001**
SNR +2 dB	-.629	.249	**< .001**
SNR -5 dB	-.936	.243	**< .001**
Intensity range	-.302	.097	**< .001**

3.4.3. Sadness

Table 13 displays the estimates of the fixed effects in the best-fitting model for the analysis of the automatic recognition of sadness (significant p-values are displayed in bold for ease of reading). Cross-lingual automatic recognition of sadness was modulated by mean F0, the slope of the LTAS and F0 variability (see Table 13). Only the slope of the LTAS was solely part of a main effect: a higher slope of the LTAS was associated with more correct sadness answers. Mean F0 was associated with AER in general and in interaction with gender (mean F0 × gender: male), as well as F0 variability (F0 variability × gender: male).

These patterns show almost no similarities with the analysis of human recognition of sadness (Scharenborg et al., 2018). Only the interaction between F0 variability and gender was found for AER and HER both. Four additional acoustic features modulated human recognition of sadness. The F0 range, mean intensity and the Hammarberg Index were all associated with human recognition of sadness, in general as well as in interaction with gender (F0 range × gender: male; mean intensity × gender: male; Hammarberg Index × gender: male). Mean F0 was

associated with human recognition of sadness in both noise conditions (Mean F0 × SNR +2 dB; mean F0 × SNR -5 dB; HER).

The presence of noise was found to have a differential influence on sadness responses from the SVM in both noise conditions compared to clean (see Table 13) (SNR +2 dB × gender: male). For HER, sadness recognition significantly deteriorated in SNR -5 dB only, compared to clean.

In sum, for sadness, F0 variability is the only feature that was found to influence both AER and HER, and only in interaction with gender. For both AER and HER, other additional features were important for correct recognition of sadness. Noise generally affects sadness in AER, as well as in interaction with gender, while noise only affects HER in at an SNR-level of -5 dB.

Table 13: fixed effect estimates for best-fitting model of the accuracy analysis of sadness (AER)

Fixed effect	**β**	***SE***	***p***
Intercept	-.267	.524	.610
Gender: male	5.858	4.427	.186
SNR +2 dB	-1.931	.576	**< .001**
SNR -5 dB	-1.316	.580	**.023**
Mean F0	-1.709	.751	**.023**
F0 variability	-.115	.766	.881
Slope of the LTAS	.559	.155	**< .001**
Hammarberg Index	.327	.284	.250
Mean F0 × gender: male	14.132	3.841	**< .001**
F0 variability × gender: male	-15.881	5.282	**.003**
SNR +2 dB × Hammarberg Index	.567	.375	.130
SNR -5 dB × Hammarberg Index	-.333	.376	.376
SNR +2 dB × gender: male	2.487	.822	**.003**
SNR -5 dB × gender: male	1.036	.824	.209

3.4.4. Joy

Table 14 displays the estimates of the fixed effects in the best-fitting model for the analysis of the automatic recognition of joy (significant p-values are displayed in bold for ease of reading).

F0 variability and the Hammarberg Index were both only part of a main effect. More variability in the F0 was generally associated with more correct joy responses from the SVM, and an increased Hammarberg Index predicted an overall decrease in correct joy responses. Furthermore, mean F0 and F0 range were associated with automatic recognition of joy in general, as well as in interaction with noise. Mean F0 only interacted with SNR -5 dB, while F0 range interacted with both noise conditions (Mean F0 × SNR -5 dB; F0 range × SNR +2 dB; F0 range × SNR -5 dB; AER). Intensity range modulated automatic recognition of joy only in interaction with gender (intensity range × gender: male).

The analysis of human recognition of joy showed that exactly the same acoustic features played a role in HER (Scharenborg et al., 2018). Mean F0, F0 range, F0 variability and the Hammarberg Index all modulated human recognition of joy. Furthermore, in background noise F0 range modulated HER at both SNR-levels (F0 range × SNR +2 dB; F0 range × SNR -5 dB), while F0 mean only did so at SNR +2 dB (mean F0 × SNR -5 dB). Intensity range modulated HER in interaction with gender (intensity range × gender: male).

Joy utterances from the male speaker were recognized differently than utterances from the female speaker in AER. This was observed for HER as well. The presence of background noise affected joy recognition in both noise conditions compared to clean, in both AER and HER.

Altogether, the acoustic features that play a role in automatic and human recognition of joy are identical. All acoustic features associated with cross-lingual AER were found to contribute similarly to cross-lingual HER. Both background noise and gender of the speakers modulate automatic and human recognition of joy.

Table 14: fixed effect estimates for best-fitting model of the accuracy analysis of joy (AER)

Fixed effect	**β**	***SE***	***p***
Intercept	-.331	.423	.435
Gender: male	5.074	1.181	**< .001**
SNR + 2 dB	-.700	.214	**.002**
SNR -5 dB	-1.884	.250	**< .001**
Mean F0	1.383	.378	**< .001**
F0 range	-1.023	.423	**.016**

F0 variability	1.290	.421	**.002**
Intensity range	-.267	.299	.372
Hammarberg Index	-.524	.169	**.002**
Mean F0 × SNR +2 dB	-.290	.287	.312
Mean F0 × SNR -5 dB	-.929	.343	**.007**
F0 range × SNR +2 dB	.816	.303	**.007**
F0 range × SNR -5 dB	1.110	.356	**.002**
Intensity range × gender: male	4.001	1.054	**< .001**

Table 15: Overview of all significant effects observed in the AER and HER per-emotion analyses concerning the role of the acoustic features, noise and gender.

Fear			**Anger**			**Sadness**			**Joy**		
Effect (on accuracy)	AER	HER	Effect (on accuracy)	AER	HER	Effect (on accuracy)	AER	HER	Effect (on accuracy)	AER	HER
Main			*Main*			*Main*			*Main*		
SNR +2 dB	↓	↓	SNR +2 dB	↓	↓	SNR +2 dB	↓	-	SNR +2 dB	↓	↓
SNR -5 dB	↓	↓	SNR -5 dB	↓	↓	SNR -5 dB	↓	↓	SNR -5 dB	↓	↓
Gender: Male	↑	↑	Gender: Male	-	↓	Gender: Male	-	-	Gender: Male	↑	↑
Mean F0	-	↑	Mean intensity	-	↓	Mean F0	↓	-	Mean F0	↑	↑
F0 range	↓	↓	Intensity range	↓	↓	F0 range	-	↑	F0 range	↓	↓
F0 variability	↑	↑	Hammarberg Index	-	↑	F0 variability	↓	-	F0 variability	↑	↑
						Slope LTAS	↑	-	Hammarberg Index	↓	↓
Interactions			*Interactions*			*Interactions*			*Interactions*		
SNR -5 dB × male	↓	↓	Mean F0 × male	-	↓	SNR +2 dB × male	↓	-	SNR -5 dB × mean F0	↓	↓
F0 mean × male	-	↑	F0 range × male	-	↓	SNR +2 dB × mean F0	-	↓	SNR +2 dB × F0 range	↑	↑
F0 variability × male	↑	↑	F0 variability × male	-	↓	SNR -5 dB × F0 mean	-	↓	SNR -5 dB × F0 range	↑	↑
			Mean intensity × male	-	↑	Mean F0 × male	↑	-	Intensity range × male	↑	↑
			Intensity range × male	-	↑	F0 range × male	-	↓			
			Slope LTAS × male	-	↓	F0 variability × male	↓	↑			
			Hammarberg Index × male	-	↓	Mean intensity × male	-	↓			
						Hammarberg Index × male	-	↓			

4. Discussion

4.1. General discussion

This study examined cross-lingual automatic emotion recognition in speech, with and without the presence of background noise. I further studied the role of seven acoustic features in automatic recognition of the investigated emotions. I compared the AER results to HER results obtained in a previous study (Scharenborg et al., 2018). This study focused on human cross-lingual acoustic emotion recognition of the same data set in the same listening conditions, and that investigated the role of the same acoustic features as in the current study, but in cross-lingual HER in noise. The main question of the current study was *how does cross-lingual AER in noise compare to cross-lingual HER in noise?* This question was split up in several smaller questions: (1) how well does a cross-lingual AER model perform in clean and noise?, (2) what is the role of the acoustic features in cross-lingual AER in noise? and (3) how does cross-lingual AER compare to cross-lingual HER (3a) in its performance and (3b) in the role of the acoustic features?

Testing of the Dutch SVM on Italian in clean and the two noise conditions yielded an overall accuracy of just above the chance-level of 25%. Investigation of the monolingual AER results with the Dutch and Italian baseline models showed much higher performance than the cross-lingual AER results from the Dutch model that was tested on Italian, which shows that in 'optimal' settings, the SVM performed well. This indicates that the Italian emotional speech data is not inherently more difficult to recognize than the Dutch emotional speech data. However, in a cross-lingual situation where the SVM was confronted with an unknown language performance drops: apparently the Dutch SVM was not well able to make the transfer from Dutch emotions to Italian emotions. The overwhelming drop in performance indicates a floor effect, where the accuracies are very low due to the SVM's inability to transfer from the Dutch emotions to the Italian emotions.

Furthermore, the accuracies obtained for clean, noise at +2 dB and noise at -5 dB were comparable and all close to chance-level. This indicates that the presence of background noise did not have a negative impact on cross-lingual AER. Most likely this was due to the aforementioned floor effect: because the accuracies were generally very low already, the presence of background noise did not cause a further decrease in the accuracies. From the

statistical analysis no effect of noise was indeed observed, however, the analysis of the role of the acoustic features and the subsequent emotion analyses did show some effects of noise. This will be discussed in more detail in section 4.6.

For recognition of anger, fear and joy I observed accuracies close to or below chance-level, while sadness was recognized very well. A confusion matrix showed that sadness was most likely the emotion the SVM 'backed off' to when a specific emotion portrayal was unclear. The finding that the SVM overwhelmingly chose sadness indicates that it experienced difficulties in recognizing the Italian data, which is most likely because of a lack of language transfer from the SVM from Dutch emotions to Italian emotions. However, I cannot exclude the possibility that (acoustic) differences between the KorEmo corpus and the EMOVO corpus that I currently could not control for also contributed to the low cross-lingual accuracy results.

The cross-lingual AER accuracies differed substantially from the cross-lingual HER accuracies in Scharenborg et al. (2018), wherein anger was the best recognized emotion, with sadness, fear and joy being recognized less well, but still well-above chance-level. Overall, the humans outperform the SVM in almost all conditions. Furthermore, the cross-lingual HER accuracies showed a linear decline for each of the four emotions from clean to noise at SNR +2 dB to noise at SNR -5 dB, compared to almost no differences between the noise conditions in cross-lingual AER.

Statistical analysis of the overall cross-lingual AER performance (i.e. the first statistical analysis described in the results) showed that in AER, sadness was recognized significantly better than the other emotions. No further effects were observed for cross-lingual AER here. The HER results from the overall analysis in Scharenborg et al. (2018) indicate that anger was recognized significantly better than all other emotions, and that emotions in the noise conditions were recognized significantly worse than in clean. The recognition patterns of cross-lingual AER and cross-lingual HER differ, and while performance of the SVM seemed unaffected by the presence of background noise, performance of the human listeners in Scharenborg et al. (2018) was affected in both noise conditions.

The acoustic features that play a role in overall cross-lingual AER in noise differ substantially from those that play a role in cross-lingual HER in noise: mean F0, F0 range, mean

intensity, intensity range, the slope of the LTAS and the Hammarberg Index were all associated with overall cross-lingual AER in noise, while only intensity range is associated with overall cross-lingual HER in noise (Scharenborg et al., 2018). This feature is thereby also the only feature that is important for both AER and HER in general. The interactions also show very few similarities between AER and HER, with the exception of overlap between three effects: F0 variability plays a role in recognition of joy for both AER and HER; the slope of the LTAS plays a role in recognition of sadness in both AER and HER; and mean F0 plays a role in emotion recognition in SNR +2 dB in both AER and HER.

To explore in more detail the role the acoustic features play in cross-lingual automatic recognition of each of the investigated emotions, separate emotion analyses were carried out and the results thereof were compared between AER and HER. These comparisons showed that the acoustic features that were important for cross-lingual recognition of anger and sadness showed considerable differences between AER and HER, while the acoustic features that played a role in recognition of fear were nearly identical in AER and HER, and the acoustic features that played a role in recognition of joy were completely identical in AER and HER. I will discuss these findings in more detail below, in separate sections.

4.2. Sadness

The cross-lingual AER results showed that sadness was recognized significantly better than the other emotions, and the accuracy for sadness was highest in all listening conditions. The high accuracy for sadness in the current study contrasts with the findings of Scharenborg et al. (2018), where sadness accuracies ranged between 60% and 30% correct depending on the listening condition. Previous studies have also observed high accuracies for both human and automatic cross-lingual recognition of sadness (Banse & Scherer, 1996; Jeon et al., 2013; Pell et al., 2009; Scherer et al., 2001; Schuller et al., 2006; Schuller et al., 2002; Thompson & Balkwill, 2006; Zhao et al., 2013). While the high accuracy for sadness is thus not necessarily surprising, in the current study it was somewhat unexpected; the human results obtained in Scharenborg et al. (2018) showed that sadness was not among the best recognized emotions, which was also observed in Giovannella et al. (2009) and (Giovannella et al., 2012).

The high accuracy in the current study is probably also due to the SVM choosing sadness every time it could not determine which emotion was expressed. Sadness was chosen most often regardless of which emotion an utterance contained. This seems to suggest that the Dutch SVM is not able to recognize the Italian emotions well but rather is 'backing off' to sadness irrespective of the emotion in the Italian utterance. Sadness thus functioned as some sort of 'I don't know'-option, which skews the accuracy observed in the current study. This emphasizes the importance of an 'I don't know'-option in emotion recognition studies with a categorical approach, which has previously been observed in HER as well (Koemans, 2016; Scharenborg et al., 2018). In the current study I chose not to add such a category, because it is not typically used in AER studies. Moreover, it is not desirable to implement an 'I don't know'-option in an automatic emotion recognition system in applications, as you would want to make sure an automatic system identifies what is heard, rather than provide it with 'an easy way out'. However, for future cross-lingual AER studies the implementation of an 'I don't know'-option might help gain a better understanding of how well an automatic system truly identifies emotions.

The acoustic feature pattern associated with cross-lingual AER of sadness might explain why it is so often confused with other emotions in the current study: a higher slope of the LTAS is uniquely associated with more correct answers for sadness, so perhaps utterances with an increased slope of the LTAS were 'automatically' recognized as being sad, thereby overruling other acoustic features. In other words, this specific feature might have predominated the presence of other features more strongly than in cross-lingual HER, resulting in the overwhelming choice for sadness in cross-lingual AER.

4.3. Anger

Anger tends to be recognized very well in AER studies, especially without background noise, though also in noise (Jeon et al., 2013; Parada-Cabaleiro et al., 2017; Schuller et al., 2002; Zhao et al., 2013), and also in HER studies (though the effect of noise is largely understudied for HER) (Koemans, 2016; Pell et al., 2009; Scharenborg et al., 2018; Scherer et al., 2001). In the current study, anger was recognized below chance-level, which is not in line with previous studies, and

more importantly, not in line with the results previously obtained for the human listeners in (Scharenborg et al., 2018).

A possible explanation for this unexpectedly low recognition rate for anger is twofold: on the one hand, the SVM's 'preference' for sadness resulted in few possible 'hits' being left for the other emotions, because of which the recognition rates could never reach high numbers anymore. On the other hand, because AER uses only intensity range and mean intensity to recognize anger, it can easily be confused with other emotions, especially if these features were not well extracted from the Italian speech. Moreover, intensity range was associated with recognition of joy as well, so it did not solely contribute to anger recognition, which probably contributed to the SVM's confusion even more. The human listeners in Scharenborg et al. (2018) on the other hand used the most features for anger recognition compared to the other emotions (seven for anger, five for sadness and joy, and three for fear). The fact that the human listeners were able to use many acoustic features for the recognition of anger might explain why anger was the best recognized emotion in Scharenborg et al. (2018), because it most likely made it easier for the human listeners to recognize anger compared to the other emotions.

4.4. Fear and joy

I will discuss the findings for fear and joy together, because their recognition and acoustic feature patterns are similar in the current study. Previously observed recognition rates of fear and joy are often lower compared to other investigated emotions (Scharenborg et al., 2018; Scherer et al., 2001; Thompson & Balkwill, 2006), especially when noise is present (Scharenborg et al., 2018; Schuller et al., 2002; Zhao et al., 2013). However, while the fear and joy accuracies are low for both AER and HER, in HER they were still well-above chance-level, compared to around chance-level for AER. Again, this indicates that the SVM was not able to transfer from the Dutch emotions to the Italian emotions.

Interestingly, the acoustic features that were found to play a role in automatic recognition of fear and joy in the current study were very similar to those that were important for the human listeners in (Scharenborg et al., 2018). All features that were found to be important for automatic recognition of fear were also important for the human listeners, who used one additional feature

compared to AER. The features that played a role joy recognition were exactly the same in AER and HER. While this did not cause the joy and fear accuracies to be comparable in AER and HER – which one might expect – it is nonetheless an interesting observation that the acoustic feature for these specific emotions patterns show so much overlap between AER and HER. Perhaps in a situation where a machine is not hindered by a language transfer (or other impeding conditions), automatic recognition of fear and joy becomes better, or more comparable to human performance. However, future research should further look into the relation between the acoustic properties of fear and joy and how they are recognized in AER and HER.

4.5. Comparing the AER results and the HER results

First of all, I compared by hand the results from the per-emotion analysis in Scharenborg et al. (2018) with the results from the per-emotion analysis in the current study, due to the data sets being of different sizes, rather than in a statistical analysis. While the observed similarities and differences are thus not statistically assessed, they are still meaningful with respect to the acoustic feature patterns observed in both studies. Nevertheless, future research should take into account that sufficient data is necessary to investigate AER, especially when focusing on cross-lingual AER and/or AER in noise, as more conditions require more data, and even more so when drawing comparisons such as the one that is provided here.

The statistical analysis where the acoustic features were added showed that the SVM overall made use of more acoustic features than the humans listeners in Scharenborg et al. (2018) did. However, the subsequent emotion analyses showed that for most emotions, the human listeners were able to make use of more features than the machines. Additionally, the recognition rates observed in the HER study were much higher than those observed in the current study. An explanation for these differences might be that the cross-lingual SVM in the current study was not fully able to use the acoustic information that was available in the speech signal, or perhaps the features used in the current study were less suitable for cross-lingual AER than they were for the human listeners in Scharenborg et al. (2018).

The fact that I used a different feature extraction script than in the previous study might explain the difference between the current AER results and the results from Scharenborg et al.

(2018). While the scripts were comparable, it might be that they did not extract the exact same information about the acoustic features. If so, the cross-lingual SVM might have had different information available than the human listeners, and as a result might have recognized the specific features differently than the humans did. However, the script still extracted the same acoustic features as in the HER study, and was found to be useful in a previous internship project carried out by me. It is therefore not very likely that the use of this feature extraction script caused large differences between the AER and HER results.

Another explanation for the finding that the AER performance was much lower than the HER performance, lies in the possibility that the acoustic feature set used in the current study was selected with the focus on HER studies. In Scharenborg et al. (2018) we investigated whether a pre-selected set of features was helpful for the human listeners. However, this pre-selected set was based on findings from previous studies with human listeners. Their contribution to HER was confirmed again in Scharenborg et al. (2018), and the AER statistical analyses indicate that they did contribute to AER in the current study. This does not exclude the possibility that using other and/or additional features might have had improved cross-lingual AER performance.

The monolingual SVM results show that the Dutch SVM performance was fine when tested on Dutch; only when tested on Italian did its performance drop. The Italian SVM that was tested on Italian performed fine as well, and did not perform worse than the monolingual Dutch SVM. This indicates that the SVM performs fine in 'standard' circumstances, and, because the Dutch and Italian monolingual SVM performances did not significantly differ from each other, it suggests that the acoustic features are at least suitable for monolingual AER in both languages. This again suggests that something went wrong in the process of transferring the acoustic information from the Dutch speech to the Italian speech. While the set of acoustic features in itself should thus not be problematic, it might still be a possibility that the set of features was too limited for cross-lingual AER, while it sufficed for the human listeners.

The differences observed between the acoustic feature patterns in the current study and in Scharenborg et al. (2018) highlights that AER and HER do not make use of the acoustic information in a speech signal in the same way, for at least two of the investigated emotions. Apparently, recognition of anger and sadness does not happen in similar ways in AER and HER,

while recognition of fear and joy does. This means that existing information from HER studies on recognition of fear and joy might be useful for improving AER, while this is most likely not the case for anger and sadness. Furthermore, the SVM suffered from the language transfer so much that noise did not further impede the SVM's performance. Information of this kind can be extremely helpful in order to optimize cross-lingual AER: now I know that AER suffers from a language transfer more than it seems to suffer from noise, I know that this is where cross-lingual AER should be improved first, for instance by providing more detailed or additional acoustic information in training a cross-lingual SVM. The human listeners on the other hand did suffer from the noise, but their recognition remained good even in the worst noise condition. Comparing the AER results to the HER results suggests that obstacles arise at different point in the recognition process for AER (the language transfer) than they do for HER (noise).

4.6. Noise

In the current study, no effect of noise on cross-lingual AER was observed, in contrast to previous studies where noise did affect AER (Parada-Cabaleiro et al., 2017; Schuller et al., 2006; Schuller et al., 2007; Zhao et al., 2013). This lack of an effect of noise is most likely due to a floor effect: because the SVM's performance was generally very low due to the SVM's inability to transfer from Dutch to Italian, accuracies did not significantly drop in the presence of background noise. On the other hand, accuracies were not 0%, so theoretically they could have dropped further in the presence of noise compared to clean. Several effects of noise were observed in the separate emotion analyses where the acoustic features were added to the analysis. The presence of background noise did affect cross-lingual AER, but how this happened precisely does not become clear from the current results.

The AER accuracies showed variation between the clean condition and the noise conditions. However, while one would expect the accuracies to drop in the presence of background noise based on previous studies, this was not observed in the current study. Rather, the overall AER accuracy observed in +2 dB was higher than in clean, which was also observed for automatic recognition of joy. Automatic recognition of anger was highest in SNR -5 dB, and automatic recognition of sadness was higher in both noise conditions than in clean. As such, AER

might be affected differently by the presence of background noise than was observed for HER in (Scharenborg et al., 2018). However, except for sadness, even the highest accuracies (divided per emotion) were all at or below chance-level, which again indicates a floor effect. Because none of the observed differences were found to be significant in the current study, future research should point out whether cross-lingual AER is indeed affected differently (or unaffected) by the presence of background noise than HER (which was negatively affected), or whether the lack of a noise effect in the current study is indeed due to a floor effect.

An interesting observation was that the SVM that was trained on a combination of Dutch and Italian did show significant improvements in performance when tested on Italian without noise, in comparison to the all-Dutch SVM tested on Italian without noise, but no differences were observed between performance of the all-Dutch and Dutch-Italian SVMs when tested in the noise conditions. For both SVM-types performance was around chance-level. Perhaps in a situation where the SVM is affected less by the language transfer, an effect of noise will show up, but future research should further look into this possibility.

Importantly, to my knowledge the current study is the first to consider the effect of babble noise on AER rather than static noise. Because babble noise is made up of speech, its acoustics are different compared to static noise types (Garcia Lecumberri et al., 2010). Therefore babble noise might affect (cross-lingual) AER differently than static noise has been observed to do, which also might be an explanation for the lack of a noise effect in the current study. However, due to the absence of a noise effect, I cannot conclude anything on this question from the results of the current study. Nonetheless, future research should look into the effect of babble noise on AER (compared to static noise).

4.7. Suggestions for future research

First of all, in the current study no effect of noise on AER was observed. However, from the current results I cannot definitively conclude whether the absence of a noise effect was due to a floor effect, or because cross-lingual AER was truly unaffected by the presence of background noise. Therefore, future studies should look into the effect of noise on cross-lingual AER, and especially considering babble noise, because this has to my knowledge not been studied before.

In order to tease apart the effect of a language transfer and the (lack of) an effect of noise, future studies should consider the effect of babble noise in monolingual situations in comparison with cross-lingual situations, and ideally also in comparison with other noise types. This would allow researchers to investigate whether AER is unaffected by babble noise; whether babble noise has a differential impact on monolingual AER than it does on cross-lingual AER; and whether babble noise has a differential impact on AER than other noise types.

Future research should further take into account that AER and HER might not make use of acoustic information in the same way, and thus should utilize other or more extensive sets of acoustic features to explore which information is most important for either AER or HER. Especially in comparative studies such as this one, it would provide interesting opportunities to compare in more detail which acoustic features humans and machines would 'choose' to use from an extensive set of features. Providing more extensive sets of acoustic features for training of automatic systems can also help improve their performance.

Finally, more emotions and languages should be considered in comparative studies that focus on both AER and HER, to investigate to what extent AER and HER are similar in how (well) they recognize specific emotions; and to study the capabilities of cross-lingual AER in more detail. The current study focused on Dutch and Italian because the data were available, but comparisons between more closely related and/or more distant languages would likely yield different results. Moreover, I only compared between four emotions, while humans know many more, and machines should eventually also be able to recognize many more. Knowledge on how humans and machines recognize emotions should thus be extended to more languages and emotions to be able to optimize AER. The best way to do so would be by creating data sets that are suitable for both AER and HER studies, which would therefore allow for more direct comparisons and ultimately more communication between the fields of automatic and human emotion recognition. However, as this may be an optimistic idea for the future, one could already start by performing more replication studies of existing AER studies with human participants, and vice versa, which allows for direct comparisons such as the one provided here.

5. Conclusion

This study provides one of the first comparisons between AER and HER on the same data set, and adds to the existing body of research on cross-lingual AER and AER in noise. It is also one of the first studies focusing on cross-lingual AER in noise, which is generally investigated separately, and is the first AER study to consider babble noise. Cross-lingual AER performance was lower compared to cross-lingual HER performance for three of the four investigated emotions and in all listening conditions. Cross-lingual AER was not affected by noise, however, because the observed AER performance was low, it is not completely clear whether the absence of a noise effect was due to a floor effect or because the babble noise used in the current study truly did not have a (negative) impact on cross-lingual AER. Moreover, separate emotion analyses did show interactions between noise and several of the emotions and/or acoustic features. The acoustic feature patterns associated with automatic recognition of anger and sadness were different than those observed in HER, while the patterns observed in automatic recognition of joy and fear were nearly identical. The findings indicate that humans outperform machines in the challenging communicative environment of cross-lingual emotion recognition in noise. However, comparing the acoustic feature patterns associated with the investigated emotions also shows some similarities between AER and HER in how they use acoustic information in a speech signal to identify emotions. This information could be key for the improvement of AER systems for the ultimate purpose of using them in commercial applications.

6. References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614-636. doi:10.1037/0022-3514.70.3.614

Banziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion, 12*(5), 1161-1179. doi:10.1037/a0025827

Barrett, L. F. (1998). Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. *Cognition & Emotion, 12*(4), 579-599. doi:10.1080/026999398379574

Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., . . . Amir, N. (2011). The Automatic Recognition of Emotions in Speech. In *Emotion-Oriented Systems* (pp. 71-99).

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer (Version 6.0.49). Retrieved from praat.org

Campbell, N. (2000). *Databases of emotional speech.* Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction* (pp. 92-103): Springer.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST), 2*(3), 1-27.

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., . . . Christensen, J. A. (2015). Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Hearing research, 322*, 151-162.

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). *Emovo corpus: an italian emotional speech database.* Paper presented at the International Conference on Language Resources and Evaluation (LREC 2014).

Ekman, P. (1992a). Are there basic emotions? *Psychological Review, 99*(3), 550-553. doi:10.1037/0033-295x.99.3.550

Ekman, P. (1992b). An Argument for Basic Emotions. *Cognition & Emotion, 6*(3-4), 169-200. doi:Doi 10.1080/02699939208411068

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion, 98*(45-60), 16.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572-587.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*(2), 203-235. doi:10.1037/0033-2909.128.2.203

Elfenbein, H. A., & Ambady, N. (2003). Universals and cultural differences in recognizing emotions. *Current directions in psychological science, 12*(5), 159-164.

Elfenbein, H. A., Martin, B., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. *Emotion, 7*(1), 131-146. doi:10.1037/1528-3542.7.1.131

Feraru, S. M., Schuller, D., & Schuller, B. (2015). *Cross-language acoustic emotion recognition: An overview and some tendencies.* Paper presented at the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII).

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychol Sci, 18*(12), 1050-1057. doi:10.1111/j.1467-9280.2007.02024.x

Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication, 52*(11-12), 864-886. doi:10.1016/j.specom.2010.08.014

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014a). Corrigendum: Cultural Relativity in Perceiving Emotion From Vocalizations. *Psychol Sci, 25*(12), 2284. doi:10.1177/0956797614556283

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014b). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion, 14*(2), 251-262. doi:10.1037/a0036052

Giovannella, C., Conflitti, D., Santoboni, R., & Paoloni, A. (2009). *Transmission of vocal emotion: Do we have to care about the listener? The case of the Italian speech corpus EMOVO.* Paper presented at the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops.

Giovannella, C., Floris, D., & Paoloni, A. (2012). An exploration on possible correlations among perception and physical characteristics of EMOVO emotional portrayals. *IxD&A, 15*, 102-111.

Goudbeek, M., & Broersma, M. (2010). *The Demo/Kemo corpus: A principled approach to the study of cross-cultural differences in the vocal expression and perception of emotion.* Paper presented at the 7th International Conference on Language Resources and Evaluation (LREC 2010).

Goudbeek, M., & Scherer, K. R. (2008). *Acoustic profiles in emotion-The GEMEP corpus.* Paper presented at the ISCA Tutorials and Research Workshop, Aalborg, Denmark.

Goudbeek, M., & Scherer, K. R. (2010). Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *J Acoust Soc Am, 128*(3), 1322-1336. doi:10.1121/1.3466853

Huang, C., Guoming, C., Hua, Y., Yongqiang, B., & Li, Z. (2013). Speech emotion recognition under white noise. *Archives of Acoustics, 38*(4), 457–463.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proc Natl Acad Sci U S A, 109*(19), 7241-7244. doi:10.1073/pnas.1200155109

Jeon, J. H., Le, D., Xia, R., & Liu, Y. (2013). *A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception.* Paper presented at the Interspeech.

Koemans, J. (2016). *Verbal Emotion Perception in Adverse Listening Conditions.* (BA Thesis). Radboud University, Nijmegen. Retrieved from https://theses.ubn.ru.nl/handle/123456789/3862

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology, 15*(2), 99-117. doi:10.1007/s10772-011-9125-1

Kowalska, M., & Wróbel, M. (2017). Basic emotions. *Encyclopedia of Personality and Individual Differences*, 1-6.

Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion, 5*(3), 277-295. doi:10.1037/1528-3542.5.3.277

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion, 19*(5), 633-653. doi:10.1080/02699930441000445

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K. (2011). Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language, 25*(1), 84-104.

Lee, C. M., Narayanan, S. S., & Pieraccini, R. (2002). *Combining acoustic and language information for emotion recognition.* Paper presented at the Seventh International Conference on Spoken Language Processing.

Luo, X., Fu, Q.-J., & Galvin III, J. J. (2007). Cochlear implants special issue article: Vocal emotion recognition by normal-hearing listeners and cochlear implant users. *Trends in amplification, 11*(4), 301-315.

Majid, A., & Levinson, S. C. (2010). WEIRD languages have misled us, too. *Behav Brain Sci, 33*(2-3), 103. doi:10.1017/S0140525X1000018X

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224.

Mesquita, B. (2001). Emotions in collectivist and individualist contexts. *Journal of Personality and Social Psychology, 80*(1), 68.

Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human nature, 1*(3), 261-289.

Parada-Cabaleiro, E., Baird, A., Batliner, A., Cummins, N., Hantke, S., & Schuller, B. W. (2017). *The Perception of Emotions in Noisified Nonsense Speech.* Paper presented at the INTERSPEECH.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing Emotions in a Foreign Language. *Journal of Nonverbal Behavior, 33*(2), 107-120. doi:10.1007/s10919-008-0065-7

Peter, C., & Beale, R. (2008). *Affect and emotion in human-computer interaction: From theory to applications* (Vol. 4868): Springer Science & Business Media.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology, 17*(3), 715.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc Natl Acad Sci U S A, 107*(6), 2408-2412. doi:10.1073/pnas.0908239106

Scharenborg, O., Kakouros, S., & Koemans, J. (2018). *The Effect of Noise on Emotion Perception in an Unknown Language.* Paper presented at the Proc. 9th International Conference on Speech Prosody 2018.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143-165. doi:10.1037/0033-2909.99.2.143

Scherer, K. R. (2000). *A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology.* Paper presented at the Sixth International Conference on Spoken Language Processing.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1-2), 227-256. doi:10.1016/s0167-6393(02)00084-5

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language, 27*(1), 40-58.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92. doi:Doi 10.1177/0022022101032001009

Scherer, K. R., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *Int J Psychol, 46*(6), 401-435. doi:10.1080/00207594.2011.626049

Scherer, K. R., Wallbott, H. G., & Summerfield, A. B. (1986). *Experiencing emotion: A cross-cultural study*. Paris, France: Editions de la Maison des Sciences de l'Homme.

Schmidt, J., Janse, E., & Scharenborg, O. (2016). Perception of emotion in conversational speech by younger and older listeners. *Frontiers in psychology, 7*, 781.

Schuller, B. (2018). Speech emotion recognition. *Communications of the ACM, 61*(5), 90-99. doi:10.1145/3129340

Schuller, B., Arsic, D., Wallhoff, F., & Rigoll, G. (2006). *Emotion recognition in the noise applying large acoustic feature sets.* Paper presented at the Proc. Speech Prosody 2006, Dresden.

Schuller, B., Lang, M., & Rigoll, G. (2002). *Automatic emotion recognition by the speech signal.* Paper presented at the Proc. of SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, Florida, USA.

Schuller, B., Maier, A., & Batliner, A. (2007). *Towards More Reality in the Recognition of Emotional Speech*. Paper presented at the Acoustics, Speech and Signal Processing.

Schuller, B., Steidl, S., & Batliner, A. (2009). *The interspeech 2009 emotion challenge.* Paper presented at the Tenth Annual Conference of the International Speech Communication Association.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009). *Acoustic emotion recognition: A benchmark comparison of performances.* Paper presented at the 2009 IEEE Workshop on Automatic Speech Recognition & Understanding.

Sobin, C., & Alpert, M. (1999). Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy. *Journal of Psycholinguistic Research, 28*(4), 347-365. doi:10.1023/a:1023237014909

Tao, J., & Tan, T. (2005). *Affective computing: A review.* Paper presented at the International Conference on Affective computing and intelligent interaction.

ten Bosch, L. (2003). Emotions, speech and the ASR framework. *Speech Communication, 40*(1-2), 213-225. doi:10.1016/s0167-6393(02)00083-3

Thompson, W. F., & Balkwill, L. L. (2006). Decoding speech prosody in five languages. *Semiotica: Journal of the International Association for Semiotic Studies/Revue de l'Association Internationale de Sémiotique, 2006*(158). doi:10.1515/sem.2006.017

Truong, & Raaijmakers, S. (2008). *Automatic recognition of spontaneous emotions in speech using acoustic and lexical features.* Paper presented at the International Workshop on Machine Learning for Multimodal Interaction.

Van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of Vocal Expressions of Emotion. *Journal of Cross-Cultural Psychology, 14*(4), 387-406. doi:10.1177/0022002183014004001

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3-28.

Wallbott, H. G., & Scherer, K. R. (1986). How universal and specific is emotional experience? Evidence from 27 countries on five continents. *Social Science Information, 25*(4), 763-795. doi:10.1177/053901886025004001

Wilting, J., Krahmer, E., & Swerts, M. (2006). *Real vs. acted emotional speech.* Paper presented at the Ninth International Conference on Spoken Language Processing.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (2006). *Emotion recognition from noisy speech.* Paper presented at the 2006 IEEE International Conference on Multimedia and Expo.

Zhao, X., Zhang, S., & Lei, B. (2013). Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications, 24*(7-8), 1539-1553. doi:10.1007/s00521-013-1377-z