Accent Matters in Perception of Voice Similarity

Mette Hjortshøj Sørensen
Aarhus University

Abstract
This study investigates how voice similarity is perceived by three different groups of listeners, namely by native listeners, by non-native listeners and by a group of listeners with no prior knowledge of the language. The study explores *whether* listeners can distinguish between voices and also *how similar* the listeners perceive the voices to be. The participants all listened to short recordings of 60 voice pairs of young male speakers speaking Danish and were asked to make a decision on whether they thought the voices sounded similar or not on a sliding scale. The results suggest that most of the listeners use the difference in fundamental frequency when deciding whether two voices sound similar or not. However, for the native listeners a change in regional accent seems to trump mean fundamental frequency as a deciding factor for judging voice similarity.

1. Introduction[1]

The speech signal carries tremendous amounts of information simultaneously. At the same time as the linguistic message is being delivered, indexical information about the speaker's identity, sex, regional origin, age, socioeconomic status, physical and emotional state is also present in the speech signal (Johnson, Westrek, Nazzi & Cutler, 2011).

[1] Parts of the findings reported in this contribution were presented at the IAFPA (International Association for Forensic Phonetics and Acoustics) Annual Conference 2010 in Trier (Sørensen 2010) and part of my doctoral research (Sørensen 2011). I am grateful for the comments and suggestions from the audience at IAFPA 2010. I would also like to thank Ocke-Schwen Bohn for the many discussions we have had about perception in general.

Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), ***A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*** (pp. 135-148). Dept. of English, School of Communication & Culture, Aarhus University.

Studies on perception of second language (L2) sounds have long discussed the influence that the first language (L1) inevitably will have on perception of the L2 sounds (e.g. Flege, 1993; Wode, 1995; Best, 1995). It is generally accepted that adult learners are language-specific perceivers, at least in the initial stages of L2 learning (e.g. Best & Tyler, 2007). That is, the adult language learners process the L2 segments by means of their L1 sound inventory. These studies focus primarily on the phoneme inventory in the second language.

The present study explores whether listeners also listen through the filter of their native language when they are asked to judge voice similarity. It is clear that in spoken language, segmental information cannot be completely disentangled from the indexical information that is also present in the speech signal at the same time as the linguistic message (Johnson et al 2011). Although what counts as being indexical information in one language may be matter of phoneme identity in other languages, i.e. this is rather language specific (e.g. Gordon & Ladefoged, 2001). For example, in Jalapa Mazatec creaky voice is used to signal the difference between /jq̰/ meaning "he swears" and /já/ meaning "tree" (Kirk, Ladefoged & Ladefoged, 1993) whereas creaky voice primarily serves as indexical information in English. Hence, there will be a certain language-specificity to what counts as indexical information as well as sound inventory. Kreiman and Gerratt (2010) suggest that the native language of a listener does affect the listener's sensitivity to voice characteristics as well as the perceptual strategy. Consequently, people are not surprisingly also more accurate when recognising voices in their native language compared to another language (Köster & Schiller, 1997).

Very little is known, however, about what listeners actually use as deciding cues or parameters when they listen to and apparently judge some voices to sound very similar and other voices to sound very different from one another. It may also differ what people actually consider as being part of the voice – whether it is laryngeal settings or whether some listeners also include supralaryngeal settings as part of their concept of 'voice quality'.

Grønnum (2005) asserts that *intonation* is the strongest marker of dialects or regional varieties in Danish, and Kristiansen, Maegaard and Pharao (2011) also found that Danish speakers primarily seem to use intonation as a cue when identifying different types of Danish regional varieties. In fact, Danish is often described as being a relatively uniform language regarding variation at the segmental level (e.g. Grønnum, 1994;

Kristiansen, 2003). Segmental variation used to be a prominent part of Danish dialects in the past, but the segmental variation has been replaced by intonation as the more salient feature in modern Danish (Gregersen & Pharao, 2016). The term 'regional accent' will be used in the current study to stress that the differences found in the data are primarily differences in intonation patterns. In a study by Gooskens (1997) examining whether English and Dutch listeners rely more on segmental variation or intonation when identifying dialects, the results suggest that intonation also seems to be more important for the identification of English dialects whereas it appears less important for the identification of Dutch dialects.

Studies on recognition of *voices* also show that listeners have a higher success rate at *remembering* and *recognising* speakers who have either relatively high or relatively low fundamental frequencies (F0) compared to speakers with a more average fundamental frequency and this goes for English (e.g. Foulkes and Barron, 2000) as well as for Danish listeners (Sørensen, 2012). This suggests that – at least English and Danish listeners – appear to rely heavily on speakers' F0 when listening to voices. Foulkes and Barron (2000) suggest that not only the mean F0 itself, but also the standard deviation (St. dv.) of F0 could have a correlation with the recognition rate in a speaker recognition test. Foulkes and Barron state that measuring the standard deviation is useful in some cases, as it enables a quantification of the F0 variation used by a speaker. According to Foulkes and Barron, speakers who are perceived as sounding monotonous most often would also have a lower standard deviation associated with their mean fundamental frequencies.

The aim of the present study is primarily to investigate whether voice similarity is perceived through the filter of the listener's native language like e.g. segments are (e.g. Flege, 1993; Best, 1995). The study examines whether native listeners, non-native listeners, and listeners with no prior knowledge of the language in question focus on the same or different acoustic cues when they are judging voice similarity. That is, do people listen to speakers in other ways when they listen to other languages compared to their own native language? The present study then extends upon some of the previous research by exploring whether listeners can discriminate between voices, but also by investigating how similar or different the listeners perceived the voices in the study to be. The focus in this study will be on the possible correlation between mean F0 and perceived similarity of voices. That is, would a small measured

difference in fundamental frequency entail a small perceived difference between voices and would a larger measured difference in fundamental frequency between two voices entail a larger perceived difference between the voices?

Assuming that listeners focus on different cues in the voices depending on their familiarity with the language spoken, this may have an effect on whether voices are judged to be similar or not. The underlying assumption of this voice perception study is that the listeners with no prior knowledge of a given language will have to listen to the voice quality in a more global (as opposed to local) manner than the native listeners would. In other words, listeners with no prior knowledge of the language in question would probably solely make use of suprasegmental features, as they would have no prerequisite for what else to listen for – whereas the native listeners may listen for both subtle segmental and suprasegmental information, e.g. regional accent, intonation or other linguistic features when they perceive and judge voice similarity between speakers.

2. Method

2.1 Stimuli

The stimuli consist of recordings of spontaneous speech from 15 young Danish male speakers between 20 and 35 years of age. The speakers' F0 varied, but speakers with any other distinctive/characteristic voice qualities, like e.g. nasal, hoarse or creaky voice were excluded from this study. Furthermore, occurrences of any other linguistic cues to regional variety, e.g. regional vocabulary or grammatical constructions that are region specific were excluded as well. 12 of the young male speakers form a relatively homogeneous group from Eastern Jutland in Denmark, all speaking Danish with a regional (but not strong) accent. There are three additions to this otherwise homogenous group of speakers, namely one young male speaker from the Northern part of Jutland in Denmark and two young male speakers from the Copenhagen area in Denmark. These voices were added to the study to test whether the listeners would react to a change in the regional accent spoken. Small samples of 3 seconds of duration were extracted from the speakers and these were then presented in pairs. In total the stimuli consisted of 60 voice pairs of 2 x 3 seconds of speech.

2.2. Listeners

Three groups of listeners participated in the study: A group of native listeners, a group of non-native listeners and a group of listeners with no prior knowledge of Danish. The first group was a group of 20 native listeners (21-40 years old) from Eastern Jutland in Denmark. The second group consisted of 20 non-native listeners with English as L1 (age 24-35 years old) who speak Danish as an L2 language at different levels of proficiency. It proved difficult to recruit participants with similar levels of proficiency in Danish, so the criteria for this group was that all the listeners had to be adult when arriving in Denmark, all of them lived in Denmark and all of them had first-hand knowledge of Danish. The third group, the listeners with no prior knowledge of Danish, were English L1 speakers (20-36 years old) from York in England and none of these speakers had any knowledge of Danish. All of the listeners from all of the groups self-reported normal hearing.

2.3. Procedure

The speech perception software 'Alvin' (Gayvert & Hillenbrand, 2003) was used and modified to suit the present study. The listeners all listened to 60 voice pairs over high quality headphones (Sennheiser HD 280 Pro) on a laptop. The 60 voice pairs were played in random order and all of the voice pairs occurred twice in order to explore whether the listeners were consistent in their judgements throughout the study. After listening to each voice pair, the listener was asked on the screen to make a decision on how similar the voices just heard were on a sliding scale going from "very different" on one end to "very similar" in the other end. The listener would then move the slider accordingly on the screen and press 'okay' and after this the next voice pair would be played automatically and so forth. Order effects were checked for as well in the study. That is, some of the voice pairs were not only played twice, but also in reverse order.

3. Results

As mentioned, previous research suggest that speaker's F0 may be one of the important features when listeners notice and remember voices (e.g. Foulkes and Barron, 2000; Sørensen, 2012). For the current voice similarity perception study it was therefore also a priority to examine whether the actual measured difference in fundamental frequency was also reflected

by the *perceived* similarity, i.e. whether there was actually a correlation between *measured* difference in fundamental frequency and the listeners' ratings of voice similarity.

The scatter plot in Figure 1 shows the difference in mean F0 between the voices in all the voice pairs measured in Hz on the X-axis compared with the perceived difference between the voices in the voice pairs on the Y-axis. Low numbers on the Y-axis correspond to a small perceived difference between the voices and higher numbers correspond to a larger perceived difference.

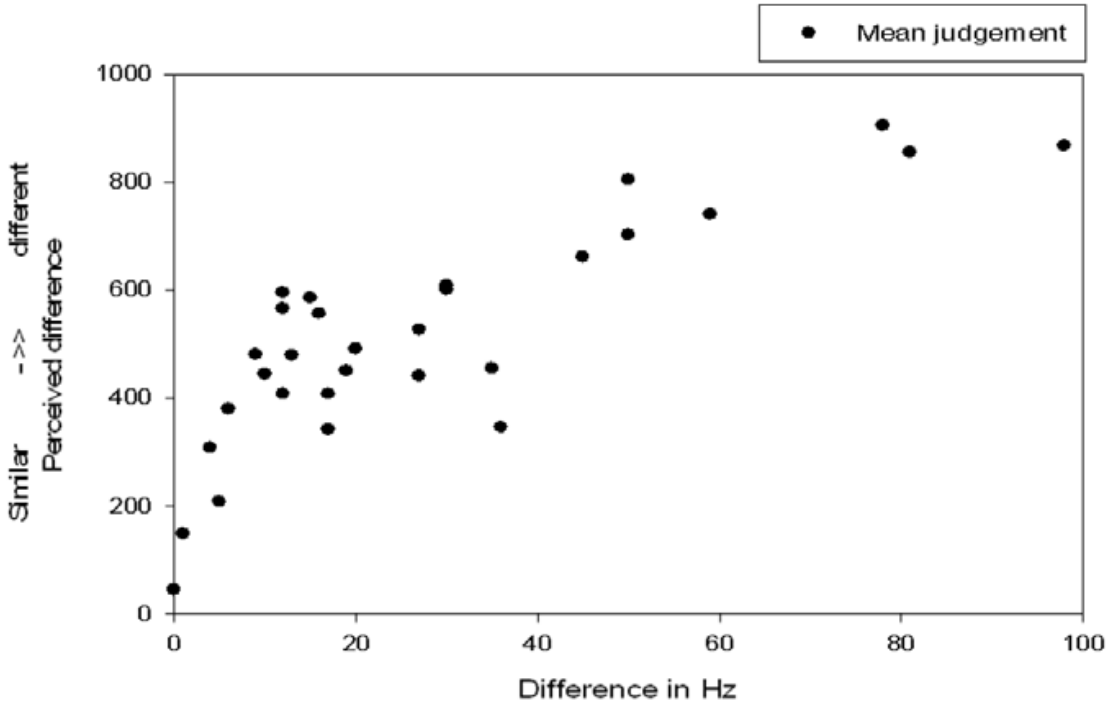

Figure 1. Results from the voice perception study showing correlation between the acoustic difference in mean F0 between the heard voices and the perceived difference between the voices.

Figure 1 shows the mean of all the listeners' trials from all the groups. The figure indicates that, in general, as the acoustic difference between the two voices in voice pair goes up, listeners will also perceive a larger difference. This was confirmed by correlation analysis (Pearson's *r*) which showed that the correlation coefficient is r=.83 (p<.001). The results from the current voice perception study suggest that, in general, most of the listeners seem to use distance – or difference – in fundamental frequency as an important cue to judge voice similarity most of the time. Figure 2

shows the same results as are shown in Figure 1, but this time the results are divided into the mean scores for each of the three different groups of listeners.

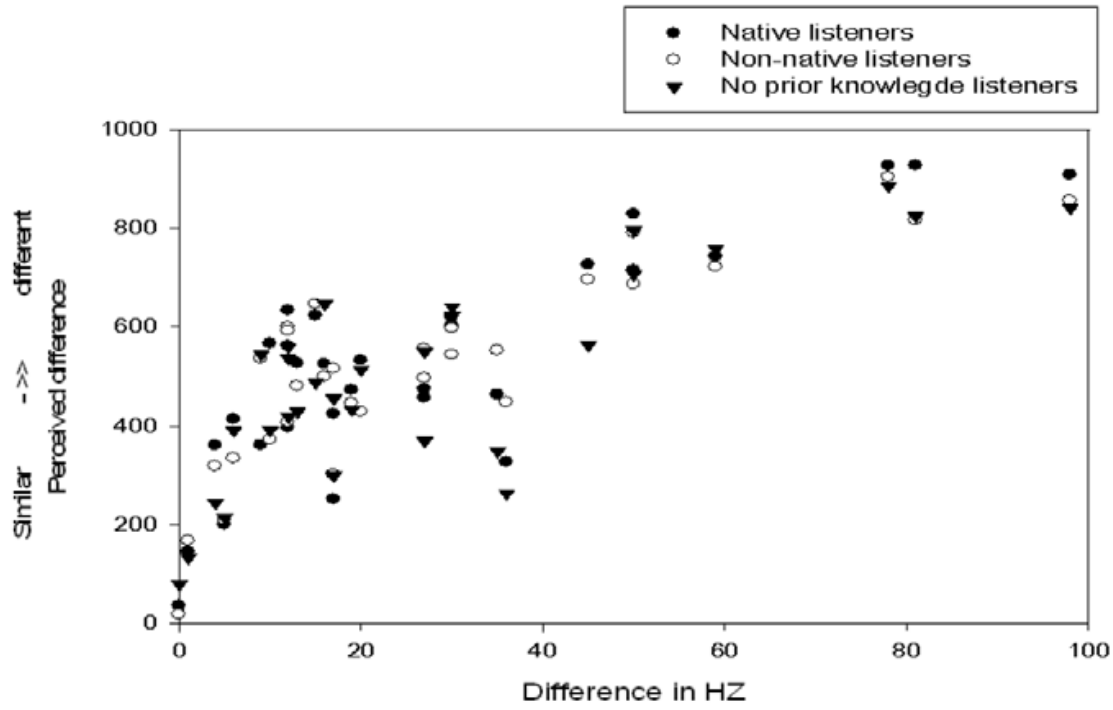

Figure 2. Results showing correlation between the measured difference in F0 between the heard voices and the perceived difference between the voices divided into the three different listener groups.

The results from Figure 2 suggest that there is a general correlation between difference between voices measured in Hz and the perceived similarity between the voices by all the three different listener groups. All three groups show a tendency to judge voices that are quite close measured in Hz to be perceptually similar. Voices that are further apart measured acoustically in Hz are generally also judged to be perceptually more different by all three groups of listeners.

The scatter plot in Figure 3 shows the mean of the first trials of all the listeners compared with the mean of all the listeners' second trial. The low numbers in the figure reflect a small perceived difference between the voice pairs and high numbers reflect a larger perceived difference. The results from the study suggest that the majority of the listeners in all three groups were consistent in their judgements from the first time to the second time they heard the same voice pair – regardless of their level of

knowledge of Danish.

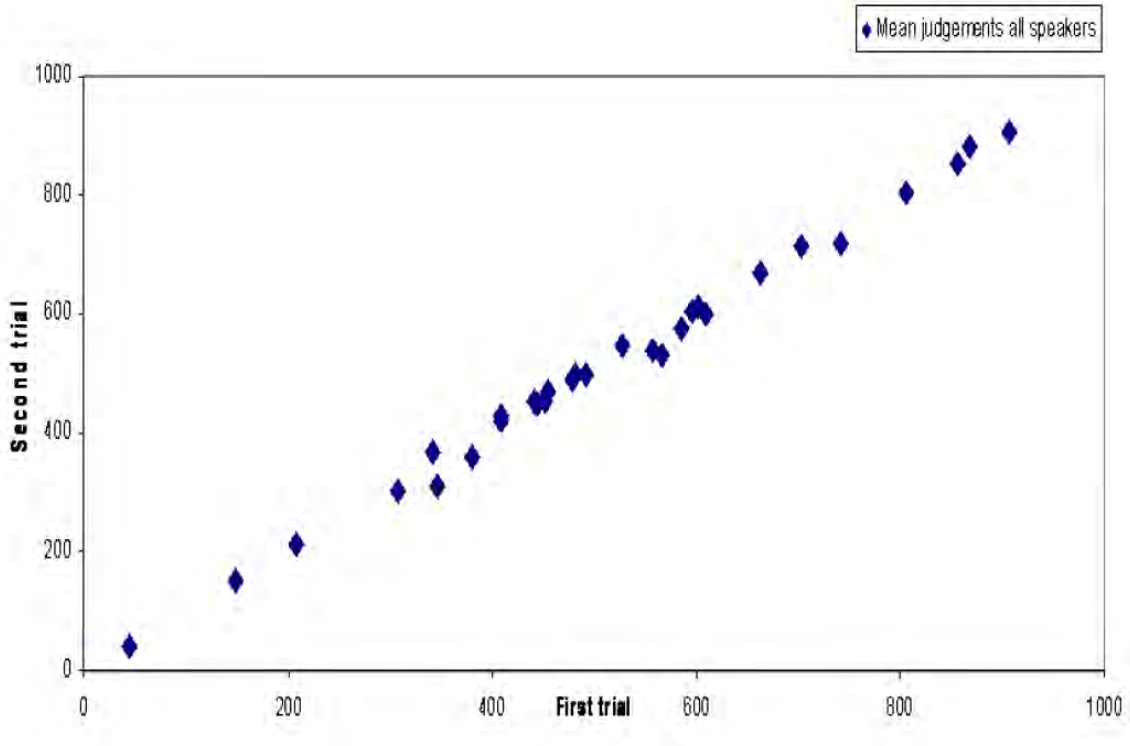

Figure 3. The mean of all the listeners' first trials correlated with the mean of all the listeners' second trial.

Figure 3 shows an almost straight diagonal line through the figure. This suggests that, generally, the listeners are consistent in their judgements from their first to their second trial. This impression was confirmed by correlation analysis (Pearson's r) which showed that the correlation coefficient is r= .97 (p<.001). In general, there appears to be a strong correlation between the acoustic difference of the mean F0 and the perceived voice similarity.

There are, however, a few exceptions to the trend of a correlation between the acoustic difference of the mean F0 and the perceived voice similarity. Figure 4 shows the results for a single voice pair where the voices were relatively similar according to fundamental frequency. There was only a measured difference of three Hz between the average fundamental frequencies for the two speakers in this sample.

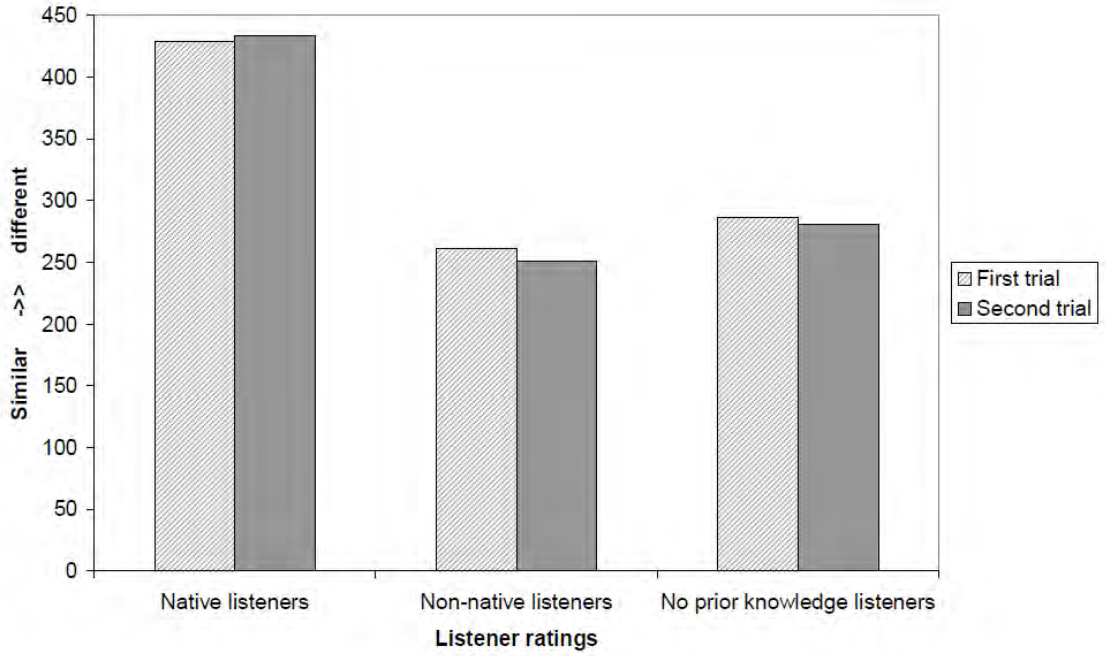

Figure 4. Results from a voice pair where the fundamental frequency is relatively similar, but the speakers speak with different regional accents.

The example in Figure 4 is particularly interesting because the two speakers in this example are from different parts of the country, namely one speaker from Eastern Jutland and the other speaker from Zealand (Copenhagen area). Apparently, the difference in regional accent between the two speakers strongly affects the way that the native listeners judge the voice pair. A one-way ANOVA was run and confirmed the visual interpretation of Figure 4 that the difference between the groups was significant, $F(2,57)=54.422$, $p=.0001$. A larger difference was perceived by the native listeners than by the two other groups.

The group with no prior knowledge of Danish would have no prerequisite for what linguistic cues to listen for whereas the native listeners could make use of language specific segmental as well as suprasegmental cues. The results from the present study showed that there were more examples similar to the one in Figure 4. This suggests that there is something in the auditory signal that the native listeners perceive which the two other groups do not when they judge voice similarity. Since the two speakers in the example have similar F0 (only a difference of 3 Hz) a

possible explanation could be that the native listeners are more sensitive to the exact intonation pattern that would be distinct for the two speakers from the two different part of the country. Another explanation could be that native listeners are listening for subtle segmental cues when judging voice similarity after all. In similar examples the results for the non-native listeners were most often closer to the ones of the listeners with no prior knowledge of Danish than they were to the native listeners. This suggests that listeners listen to voices through their L1 filter and possibly not as sensitive to exact intonation patterns or subtle segmental variation in their L2.

The results from the study suggest that, as long as it is a homogenous group of speakers, then the native listeners seem to base their judgement of voice similarity on differences in mean fundamental frequency. However, a difference in regional accent seemed to trump mean fundamental frequency for the native listeners, making some voice pairs perceived to be more different from one another than the other two groups perceived them to be.

4. Discussion

In general, the listeners seem to judge voice similarity according to fundamental frequency – at least when the voice quality of the speakers are not very distinct, such as e.g. nasal, creaky or hoarse. However, for the native listeners this seemed to be the case only when speakers spoke with the same regional accent. When there was a change in accent, this affected the perceived difference and distance between the voices. Therefore it is important to keep in mind that language specific cues play a role for native listeners, whereas listeners with no prior knowledge of a given language listen in a more global manner and that non-native listeners resemble listeners with no prior knowledge more than they resemble native listeners.

The results from the present study suggest that, in general, as the measured difference between the standard deviation of the two voices in the voice pairs goes up it is also perceived as a bigger difference by the listeners (r=.624007, p<0.01). The results suggest that there is also some correlation between difference in the mean F0 *variation* and the perceived similarity between the voices by the different listener groups which is in line with suggestions made in previous studies, e.g. Foulkes & Barron (2000). The listeners show a tendency to judge voices with similar standard deviation measured in Hz to be perceptually similar as well. Voices that differ with more F0 variation are generally also judged to be perceptually

more different.

Several studies suggest that – besides fundamental frequency – average formant frequencies over longer stretches of speech also play a part when recognising voices (e.g. Nolan and Grigoras, 2005, Jessen 2008). Even though the first three formants are related to vowel quality produced, and hence have some constraints, there are still individual speaker differences in vowel articulation (Johnson, 2003). Not only are formant frequencies essential correlates of distinctions between different vowels and some consonants, but they also convey important speaker specific information (Jessen, 2008). As formant location depends on vocal tract characteristics, e.g. longer vocal tracts generally lead to lower formant frequencies, it is also possible that the formant frequencies can reveal important speaker specific pathological or habitual features in speech, e.g. a tendency to retract the tongue or a tendency to protrude the lips while speaking. It was beyond the scope of the present study to attempt assessing how this may influence the listeners rating of voices besides fundamental frequency, but there is of course a possibility that this could also be one of the features that the listeners used to decide voice similarity in the present study.

The results suggest that a change in regional accent make the native speakers judge the voice similarity to be more different as well. As mentioned in the introduction there may be different opinions of what constitutes 'voice quality' (Köster et al., 2007), hence, also whether some voices are similar or not. Some people could listen for laryngeal characteristics and others could also include articulatory setting as part of their concept of voice quality. In the current study, a change in regional accent caused native listeners to rate the voices to be more different than the other two groups. However, whether the native listeners are listening for specific intonation pattern of the regional accents or whether they are focusing on subtle segmental differences between the accents cannot be determined from the present results. It is still intriguing that a change in regional accent results in a much larger perceived difference between the voices than for other voice pairs with the same difference in F0 between the voices.

Kreiman and Gerratt (2010) also suggest that listeners may have individual listening strategies and that these strategies may be listening for different cues. However, if this was the case in the current voice perception study, much more random results across the listener groups would have been expected. The results from this study suggest that judging

voice similarity is not just a task that is particularly challenging for any of the groups, leading to inconsistent results. The shift in perceived voice similarity appeared instantly and consistently for the native speakers when there was a change in accent whereas the two other groups consistently rated the voices to be more similar in these instances.

5. Conclusion

The aim of the present study was primarily to investigate whether voice similarity is perceived through the filter of the listener's native language like e.g. segments are (e.g. Flege, 1993; Best, 1995). Therefore the study focused on perceived voice similarity between presented voice pairs by different groups of listeners, namely by native listeners, by L2 listeners and by a group of listeners with no prior knowledge of the language.

The study furthermore explored how similar the listeners perceive the voices to be and the results from the study suggest that the majority of listeners use fundamental frequency as a key feature when rating how similar the voices sounded. When the regional accent remained the same, all three listener groups rated voice pairs with similar fundamental frequency to be similar and when there was a larger acoustic difference in fundamental frequency between the voices, the listeners also rated them as very different.

However, a few voices with different regional accent were also among the presented voice pairs in order to explore the affect that a change in accent would have on perceived voice similarity. The two non-native groups still rated voice pairs with similar fundamental frequency to be similar as before. The native group, however, noticed the change in accent and rated the voices as a lot more dissimilar and seems to trump fundamental frequency as the deciding factor when rating voice similarity. It is important to keep in mind that language specific cues play a role for native listeners, whereas listeners with no prior knowledge of a given language listen in a more global manner and that non-native listeners resemble listeners with no prior knowledge more than they resemble native listeners. The results suggest that listeners do actually listen through the filter of their native language – that this is not limited to sound inventory, but also applies when rating voice similarity. The findings from this study could have practical implications for several areas of applied phonetics.

References

Best, C. T. (1995). A direct realistic view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171-206). Timonium, MD: York Press.

Best, C. T. & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In: Bohn, Ocke-Schwen and Murray J. Munro (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–34). Amsterdam: John Benjamins.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *Journal of the Acoustical Society of America, 93*(3), 1589-1608.

Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-272). Timonium, MD: York Press.

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic linguistics*, *7*, 180-198.

Gayvert, R. T. & Hillenbrand, J. M. (2003). Open-source software for speech perception research. *The Journal of the Acoustical Society of America, 113*(4), 2260-2260.

Gooskens, C. S. (1997). On the role of prosodic and verbal information in the perception of Dutch and English language varieties. [Sl: sn].

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383-406.

Gregersen, F., & Pharao, N. (2016). Lects are perceptually invariant, productively variable: A coherent claim about Danish lects. *Lingua*, *172*, 26-44.

Grønnum, N. (1994). Rhythm, duration and pitch in regional variants of standard Danish. *Acta Linguistica Hafniensia, 27*(1), 189-218.

Grønnum, N. (2005). Fonetik og Fonologi, 3. udg. København: Akademisk Forlag.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass, 2*(4), 671-711.

Johnson, E. K., Westrek, E., Nazzi, T. & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science, 14*(5), 1002-1011.

Kirk, P. L., Ladefoged, J. & Ladefoged, P. (1993). Quantifying acoustic properties of modal, breathy, and creaky vowels in Jalapa Mazatec. *American Indian linguistics and ethnography in honor of Laurence C. Thompson*, 435-450.

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. Forensic Linguistics. *The International*

Journal of Speech, Language and the Law, 4(1).

Köster, O., Jessen, M., Khairi, F., & Eckert, H. (2007, August). Auditory-perceptual identification of voice quality by expert and non-expert listeners. In *Proceedings of the XVI International Congress of the Phonetic Sciences* (ICPhS) (pp. 1845-1848).

Kreiman, J. & Gerratt, B. R. (2010). Effects of native language on perception of voice quality. *Journal of phonetics, 38*(4), 588-593.

Kristiansen, T. (2003). Sproglig regionalisering i Danmark?. In *Nordisk dialektologi* (pp. 115-149). Novus forlag.

Kristiansen, T., Maegaard, M., & Pharao, N. (2011). Det er intonationen, vi hører det på: Perceptionsstudier i genkendelse af moderne dansk med henholdsvis jysk og københavnsk aksang. In *Jysk, Ømål, Rigsdansk Mv*, (pp. 207-224). Peter Skautrup Centret for Jysk Dialektforskning.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech Language and the Law, 12*(2), 143.

Sørensen, M. H. (2010). Perception of voice similarity by different groups of listeners, *IAFPA 19th Annual Conference*, Trier, Germany.

Sørensen, M. H. (2011). *Acoustic and perceptual aspects of speaker-specific differences in speech and their forensic implications*. Aarhus University. Doctoral thesis.

Sørensen, M. H. (2012). Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language & the Law, 19*(2).

Wode, H. (1995). Speech perception, language acquisition and linguistic: Some mutual implications. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp.321-350). Timonium, MD: York Press.