Computational Modelling of the Recognition of Foreign-Accented Speech

Odette Scharenborg[1,2], *Marijt Witteman*[1,3], and *Andrea Weber*[1,2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, the Netherlands
[3] Radboud University Nijmegen, The Netherlands
{Odette.Scharenborg, Marijt.Witteman, Andrea.Weber}@mpi.nl

Abstract

In foreign-accented speech, pronunciation typically deviates from the canonical form to some degree. For native listeners, it has been shown that word recognition is more difficult for strongly-accented words than for less strongly-accented words. Furthermore recognition of strongly-accented words becomes easier with additional exposure to the foreign accent. In this paper, listeners' behaviour was simulated with Fine-tracker, a computational model of word recognition that uses real speech as input. The simulations showed that, in line with human listeners, 1) Fine-Tracker's recognition outcome is modulated by the degree of accentedness and 2) it improves slightly after brief exposure with the accent. On the level of individual words, however, Fine-tracker failed to correctly simulate listeners' behaviour, possibly due to differences in overall familiarity with the chosen accent (German-accented Dutch) between human listeners and Fine-Tracker.

Index Terms: foreign-accented speech, accent strength, word recognition, computational modelling, German-accented Dutch

1. Introduction

Understanding spoken language seems to be one of the easiest things we do. Listeners usually handle the enormous variability in the speech signal without ever so much as noticing it. Yet this ease can diminish when listening to foreign-accented speech. In this case, the complex processes of comprehension can be easily obstructed. In foreign-accented speech, listeners are confronted with a speech signal that typically deviates noticeably from the canonical form in the target language and often reflects language-specific structures from the speaker's native language. Understanding thus requires the speech system to adapt to non-native pronunciation variations that often disagree with the structure of the target language.

Recent research with native listeners has shown that comprehension of foreign-accented speech varies for different accents and speakers, and can improve with additional listening experience (e.g., [1],[2]). Typically, these studies are interested in global foreign accents rather than specific accent markers. Specific segmental accent markers were investigated by Witteman and colleagues [3],[4] in a series of cross-modal priming studies with Dutch participants listening to German-accented Dutch. The word stimuli in Witteman et al.'s studies contained different segmental substitutions that varied in the strength of perceived accentedness. Their results show that for listeners with limited prior experience with the accent, word recognition is more difficult for strongly-accented words than for medium- or weakly-accented words. Furthermore, with very little additional exposure to the German speaker, recognition of strongly-accented words improves significantly. In the present paper, we will simulate the human ability to correctly recognise words with varying degrees of accentedness with Fine-Tracker, a computational model of spoken-word recognition.

The motivation for this study is two-fold. First, an important issue in explaining differences in recognising foreign-accented speech is teasing apart how much recognition ease is influenced by perceptual similarity between the L2 target language and the speaker's native language, or by experience with accented speech in general, or experience with a particular type of accent. Since it is basically impossible to find listeners that have never been exposed to accented speech, a truly 'monolingual' computational model might be helpful in resolving the debate. Second, as far as we know, no computational model exists that is able to simulate how human listeners recognise foreign-accented speech. Computational models have mostly focussed on explaining the recognition of unaccented speech, and are therefore typically tested on how well they handle canonical speech. Furthermore, if they are tested on non-canonical forms, the mispronunciations are set by the "experimenter" in the abstract input (e.g., [5],[6]). In this study, we will investigate the ability of an existing computational model [7] to simulate human listeners' recognition of German-accented Dutch using real speech as input.

2. Priming in German-accented Dutch for Dutch listeners

Witteman and colleagues [3],[4] investigated how both long- and short-term experience with German-accented Dutch influence word recognition by native Dutch listeners. Accented words in their study contained diphthong substitutions typical for German speakers of Dutch that either deviated acoustically from the canonical form to a large extent (*huis* [hœys], 'house', pronounced as [hɔɪs]) or to a medium extent (*lijst* [lɛɪst], 'frame', pronounced as [laɪst]). As a control, words without obvious segmental deviations were chosen (e.g., *dekking* [dɛkɪŋ], 'cover'). The mispronunciations were produced spontaneously. Varying degrees of accent strength (strong accent for [œy] words, medium accent for [ɛɪ] words, and weak accent for words with no substitutions) were confirmed in a rating study. Dutch participants with limited experience with the German accent

listened to the German-accented prime words, and subsequently made lexical decisions to printed Dutch target words. Significant facilitatory priming effects (i.e., a difference in reaction times to target words preceded by identical primes versus unrelated primes) were interpreted as successful word recognition.

Participants with limited experience with the German accent showed significant facilitatory priming for weakly-accented and for medium-accented words, but not for strongly-accented words (see Figure 1). Furthermore, the size of the priming effect was significantly smaller for the strongly-accented words than for the medium- and weakly-accented words, but medium- and weakly-accented words primed equally well. When Dutch participants first listened to the German speaker reading a short Dutch story containing words with [œy] before the cross-modal priming experiment, all word types showed significant facilitatory priming (see Figure 2).

Thus, brief additional exposure to the German-accented speaker was sufficient to immediately interpret strongly-accented words correctly in the priming study. With brief additional exposure, priming effects were furthermore comparable for the three accent types. Thus, all accent types primed equally well after exposure, although they varied in perceived accent strength. The fact that differences in perceived accentedness were not fully reflected in measurable differences in priming were attributed to 1) Dutch listeners' limited experience with German-accented Dutch (all Dutch listeners have some experience with the German accent) and 2) the close perceptual similarity of the German-accented pronunciation with the Dutch canonical pronunciation. This allowed them to recognise medium- and weakly-accented words equally well, while for strongly-accented words additional exposure was necessary to achieve immediate correct recognition.

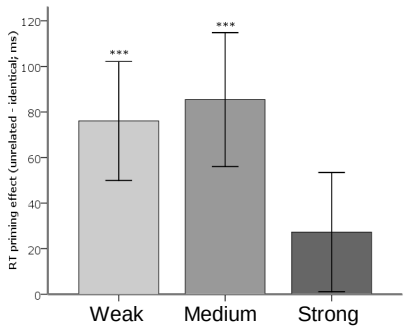

Figure 1. Priming effects for Dutch listeners with limited prior experience.

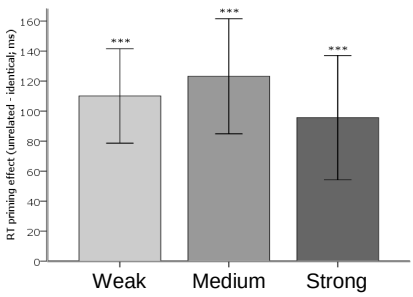

Figure 2. Priming effects for Dutch listeners with additional brief exposure.

2.1. Fine-Tracker

Fine-Tracker [7] is a computational model of human spoken-word recognition specifically developed to account for the accumulating evidence that phonetic detail is important in word recognition (e.g., [5]). It is one of only a few computational models that take the actual acoustic signal as input. The Fine-Tracker software is implemented in JAVA and is distributed via http://www.finetracker.org. In line with for example [5], Fine-Tracker assumes that the speech recognition process consists of a prelexical level and a lexical level. First, listeners map the incoming acoustic signal onto so-called prelexical representations. At the lexical level, all representations are stored in the form of sequences of prelexical units, and lexical representations that (partly) match the prelexical representations are activated in parallel.

2.2. The prelexical level

At the prelexical level, which is implemented as a set of artificial neural networks (ANNs), the acoustic signal is converted into 'articulatory feature' (AF) vectors, created for every 5 ms. AFs describe acoustic correlates of articulatory properties of speech sounds and can be used to represent the acoustic signal in a compact manner (e.g., manner and place of articulation, voicing, and tongue position during the production of vowels). The use of AFs as prelexical representations allows Fine-Tracker to 'track' and model phonetic detail in the speech signal. For more details about the AFs used in Fine-tracker see [7]. Note that one of the AFs specifically models diphthongs.

For each of the AFs, one ANN was trained for all its AF types using NICO [9]. The ANNs were trained on 3,410 randomly selected utterances from the manually transcribed read speech part of the Spoken Dutch Corpus (CGN; [10]). Fine-Tracker was only trained on speech from native Dutch professional speakers (without clear dialect markers) and can thus be regarded as a 'monolingual', i.e., a hypothetical Dutch listener who has never encountered German-accented Dutch.

For each 5 ms input frame, each ANN creates a continuous value between 0 and 1, for each of its AF types. This value can be regarded as a measure of activation of this AF type. Per input frame, all AF values are combined into a feature vector, whose length is equal to the total number of AF types. These feature vectors serve as the input of the lexical level of Fine-Tracker.

2.3. The lexical level

In Fine-Tracker's lexicon, words are also represented in terms of AF vectors. These are obtained by automatically substituting all phonemes in the lexical representations with their canonical (0 and 1) AF values. Fine-Tracker's word recognition module uses a probabilistic word search (based on Viterbi search, a standard technique in automatic speech recognition) to match the prelexical feature vectors onto the candidate words in the lexicon in order to find the most likely sequence of words. For each of the prelexical vectors the 'degree of fit' with the lexical vector is calculated, a worse fit results in a lower 'activation' of that word and vice versa. The output of Fine-Tracker is a ranked N-best list of the (in this study) 50 most likely lexical paths with likelihoods for each word on each path (the N can be set at any value). This N-best list can be created for every 5 ms time slice.

A strength of Fine-Tracker is that it can be tested with real speech rather than an abstract form of input representation as is used by other models of word recognition (e.g., [5],[6]). This

allows us to use the actual German-accented Dutch stimuli from the cross-modal priming study in [3],[4] for our simulations.

3. Set-up of the simulation

Fine-Tracker was tested in two conditions that approximate the two listener groups of the cross-modal priming study in [3],[4]: as a Dutch listener with no experience with German-accented Dutch (referred to as the *inexperienced* condition) and as a Dutch listener who had just listened to some German-accented Dutch (the *exposure* condition). Note that while Fine-Tracker truly had no prior experience with German-accented Dutch, Dutch listeners in [3],[4] had limited prior experience with the German accent. The task set to Fine-Tracker was to correctly model the varying word recognition ease observed in [3],[4]. For Fine-Tracker, we consider a target word appearing at N=1 (first best) as a correctly recognised word, while for the human data we regard a significant positive priming effect as a correctly recognised word.

For any automatic speech recognition system to work, parameters need to be set for the task and speech at hand. For the *inexperienced* condition, 94 weakly-accented filler words from the German speaker of the priming study (disjoint from the test set), with no /œy/ and /ɛɪ/ were used as a development set to tune the parameters. To simulate brief exposure to German-accented Dutch, the parameter settings for the *exposure* condition were tuned on the development set with an additional six /œy/ and six /ɛɪ/ filler words for the *inexperienced* condition. Note that the use of accented items during parameter tuning as an implementation of brief exposure is a modelling assumption, not a theoretical assumption [11].

After the optimal parameter settings were found, Fine-Tracker was evaluated using the full set of experimental words from the priming study (24 weakly-accented words, 12 medium-accented /ɛɪ/ words, and 12 strongly-accented /œy/ words). For a simulation to be considered successful, Fine-Tracker should show more correctly recognised target words with decreasing accentedness, i.e., the fewest correctly recognised words for the strongly-accented words, and more for the medium- and weakly-accented words. Secondly, we would expect that with increasing accentedness, the depth in the N-best list at which the target word was found should decrease (note: the higher the N, the worse). These are the criteria with which Fine-Tracker will be evaluated as a macroscopic computational model of the recognition of foreign-accented speech. Moreover, we will further investigate Fine-Tracker as a microscopic computational model by correlating the average depth at which a target word was found with the size of the priming effect for individual words in the priming study.

For a successful simulation of the behaviour of Dutch listeners with some exposure to German-accented Dutch, we expect Fine-Tracker's performance to increase (i.e., to recognise more words correctly) in the *exposure* condition. Moreover, we expect that with increasing accentedness, the depth in the N-best list at which the target word is found should decrease.

The Dutch lexicon used in the simulations consisted of 27,740 entries. To guide Fine-Tracker's word search, we applied priors (i.e., a higher probability) to the words in the test and development sets such that they had a higher likelihood than the other words in the lexicon. (Note, that the use of priors, usually in the form of word frequency, is standard in automatic speech recognition systems.)

Table 1. The average depth at which the target word was found in the N-best list, the average depth excluding the recognised target words, and the percentage of target words that were recognised, per accent type and per condition.

	weakly-accented targets	medium-accented targets	strongly-accented targets
inexperienced condition			
average depth	1.8	3.6	5.2
average depth, excl. first best	3.7	6.2	7.3
% target on N=1	71.4	50	33.3
exposure condition			
average depth	1.6	3.4	5.8
average depth, excl. first best	3.6	6.8	9.1
% target on N=1	76.2	58.3	41.7

Leading silences in the words of the test and development sets were cut before the stimuli were parameterized with 12 MFCC coefficients and log energy and augmented with first and second derivatives resulting in a 39-dimensional feature vector. The features were computed using 25 ms windows shifted by 5 ms per frame. The MFCC feature vectors were used as input to the ANN module at the prelexical level. The output of the prelexical level was then used as input to the search module at the lexical level of Fine-Tracker.

4. Results

In the priming study [3],[4], three weakly-accented words had been removed from the analyses due to high error rates; the same words were now removed for the analyses of the Fine-Tracker outcome. We first identified the number of target words present in the 50-best list output by Fine-Tracker. For both the *inexperienced* condition and the *exposure* condition, all 45 remaining target words appeared in the 50-best list output by Fine-Tracker. For the *inexperienced* condition a total of 25 target words were ranked first best – the lowest ranking target word was at N=15, for the *exposure* condition it was 28 words, with the lowest ranking target word at N=23. The (weighted) average depth in the N-best list at which the target words were found was 3.5 for the *inexperienced* and 3.6 for the *exposure* conditions.

To further investigate whether Fine-Tracker correctly simulated the differences human listeners showed in recognising words with varying strengths of accent, weakly-, medium-, and strongly-accented words were analysed separately. Table 1 lists the average depth at which a target word was found in the N-best list, the average depth excluding the first best target words, and the percentage of target words that were recognised, separately for each accent type and for each condition.

In line with the human data, the average depth at which a target word was found in the *inexperienced* condition was significantly higher in paired two-tailed *t*-tests for strongly-accented words than for weakly-accented words ($t(31) = 7.63$, $p < .05$), with no significant difference between medium- and weakly-accented words ($t(31) = 3.083$, $p > .05$). The number of times that the target word appeared as first best was 15 out of 21 (71.4%) for the weakly-accented words, 6 out of 12 (50%) for the medium-accented words, and only 4 out of 12 (33.3%) for the strongly-accented words.

Also in line with the human data, in the *exposure* condition there was no difference in depth of N between strongly-accented words and medium-accented words ($t < 1$) and between medium-

and weakly-accented words ($t(31) = 2.73$, $p > .1$), but contrary to the human data strongly-accented words were ranked significantly lower than weakly-accented words ($t(31) = 6.24$, $p < .05$). In the *exposure* condition, 16 (76.2%) of the weakly-accented words were ranked first best, 7 (58.3%) of the medium-accented words, and 5 (41.7%) of the strongly-accented words. Note that the number of target words ranked as first best was overall higher in the *exposure* condition than in the *inexperienced* condition, but this difference was not statistically significant. The fact that Fine-Tracker still judged strongly-accented words as less likely target words in the *exposure* condition, while human listeners recognised these words in the *exposure* condition as easily as the other words, possibly speaks for the fact that the limited previous exposure human had with the German accent helped them to adapt to strongly-accented words more quickly. Thus, we predict that with more exposure Fine-Tracker would judge strongly-accented words as more likely target words. The fact that in the *inexperienced* condition there was no significant difference between medium- and weakly-accented word for the 'monolingual' Fine-tracker, speaks for an influence of perceptual similarity on recognition ease.

For both the priming data and Fine-Tracker, there was variation in the results on an item-by-item basis, i.e., priming effects varied in size per word and for Fine-Tracker the depth at which target words were found varied. To investigate whether Fine-Tracker and the human listeners showed similar behaviour for individual words, the depth at which the target word was found in the N-best list for the *inexperienced* and the *exposure* conditions was correlated with the priming effects in the human data. A one-tailed (bivariate) Spearman's rho test of the *inexperienced* condition found no significant correlation with the inexperienced human listener group (Spearman's rho = .152, p = .159). And also, for the *exposure* condition, no significant correlation was found with the human listeners in the exposure group (Spearman's rho = -.28, p = .427). Thus, Fine-Tracker could not successfully simulate listeners' behaviour in the priming study on an item-by-item basis.

5. Summary and conclusions

Fine-Tracker was able to correctly predict the difficulty inexperienced listeners have with German-accented Dutch: in line with the human data, the average depth at which a target word was found was significantly higher for strongly-accented words than for weakly-accented words, with no significant difference between medium- and weakly-accented words. Returning to the perceptual similarity vs. experience debate, this latter result speaks for an influence of perceptual similarity on recognition ease.

The overall results improve slightly (but not significantly so) when Fine-Tracker was exposed to a few accented items during parameter tuning. Again, with increased accent, the number of recognised words decreased. Moreover, there was no difference in N-depth between strongly-accented words and medium-accented words or between medium- and weakly-accented words, but contrary to the human data strongly-accented words were ranked significantly lower than weakly-accented words.

This discrepancy between the human listeners and Fine-Tracker in the *exposure* condition can either be explained by an adaptation advantage human listeners had through prior limited exposure or it can be explained by the difference in the way human listeners and Fine-Tracker were exposed to German-accented Dutch. In Fine-Tracker, exposure was implemented as a change in parameter settings, these (and other, not reported) results indicate that this may not the optimal implementation of 'exposure'. Follow-up research will focus on incorporating exposure to accented speech at the prelexical level, i.e., by training Fine-Tracker on German-accented Dutch, such that the AFs will be adapted to German-accented Dutch. Since listeners in the priming study had limited exposure to German-accented Dutch, unlike Fine-Tracker, it is possible that the *inexperienced* condition does reflect the results as one would expect for a Dutch speaker who has never been exposed to German-accented Dutch. This interpretation needs to be investigated further.

The item-by-item analysis showed that Fine-Tracker was not able to correctly simulate listener results on an item-by-item basis. However, it should be noted that other aspects of words (such as lexical frequency) also influence recognition ease, and it is possible that differences in performance on an item-by-item basis mainly reflect differences in how Fine-tracker and human listeners react to these additional aspects.

Concluding, Fine-Tracker was able to simulate human listeners' recognition of German-accented Dutch, although more research is needed to determine the status of the model as a true monolingual of Dutch. Nevertheless, this study shows that computational modelling is a valuable asset in investigating the mechanisms underlying the recognition of foreign-accented speech.

6. Acknowledgements

The research by the first author is sponsored by the Max Planck International Research Network on Aging, Rostock, Germany. Research by the second and third author is funded by the Max Planck Society, Munich, Germany.

7. References

[1] Bradlow, A.R., Bent, T. "Perceptual adaptation to non-native speech", Cognition 106(2):707-729, 2008.

[2] Clarke, C.M., Garrett, M. "Rapid adaptation to foreign accented speech", JASA 116(6):3647-3658, 2004.

[3] Witteman, M.J., Weber, A., McQueen, J.M., "Strength of a foreign accent and listener familiarity with it co-determine speed of perceptual adaptation" (under revision).

[4] Witteman, M.J., Weber, A., McQueen, J.M., "Rapid and long-lasting adaptation to foreign-accented speech", JASA 128(4): 2486, 2010.

[5] Norris, D., "Shortlist: A connectionist model of continuous speech recognition", Cognition 52, 189-234, 1994.

[6] McClelland J.L., Elman J.L. "The TRACE model of speech perception", Cognitive Psychology 18:1-86, 1986.

[7] Scharenborg, O., "Modeling the use of durational information in human spoken-word recognition", JASA 127(6):3758-3770, 2010.

[8] Salverda, A.P., Dahan, D., McQueen, J.M., "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension", Cognition 90, 51-89, 2003.

[9] Ström, N., "Phoneme probability estimation with dynamic sparsely connected artificial neural networks", Free Speech Journal 5, 1997.

[10] Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., Baayen, H., "Experiences from the Spoken Dutch Corpus project", Proc. LREC, Las Palmas, Gran Canaria, pp. 340-347, 2002.

[11] Scharenborg, O., Boves, L. "Computational modelling of spoken-word recognition processes: design choices and evaluation", Pragmatics & Cognition 18(1):136-164, 2010.