Tilburg University

Prosodic evaluation of accent distributions in spoken news bulletins of Flemish newsreaders

Swerts, M.G.J.; Marsi, E.C.

Published in:
Journal of the Acoustical Society of America

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Swerts, M. G. J., & Marsi, E. C. (2012). Prosodic evaluation of accent distributions in spoken news bulletins of Flemish newsreaders. *Journal of the Acoustical Society of America*, 132(4), 2616-2624.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Prosodic evaluation of accent distributions in spoken news bulletins of Flemish newsreaders

Marc Swerts^{a)}

Tilburg University, Tilburg center for Cognition and Communication (TiCC), P.O. Box 90153,
NL-5000 LE Tilburg, The Netherlands

Erwin Marsi

Department of Computer and Information Science, Norwegian University of Science and Technology,
Sem Sælandsvei 7-9, NO-7491 Trondheim, Norway

(Received 18 October 2011; revised 7 June 2012; accepted 27 August 2012)

The current article describes research on whether the goodness of a particular speaking style correlates with the way speakers distribute pitch accents in their speech. Study 1 analyzed two Flemish newsreaders, who, according to poll ratings, had previously been judged to represent a good vs bad speaker. A perception study in which participants had to assess the quality of spoken paragraphs produced by either of the two speakers confirmed that one speaker was rated as significantly and consistently better than the other one. An exploration of the accent distributions in those paragraphs showed that the accent distributions of the better speaker were more similar to the ones of a gold standard, i.e., the accent distributions as predicted by two independent intonation experts. Study 2 compared synthetic versions of a selection of the paragraphs of study 1, generated by a Dutch text-to-speech system. It compared three basically identical versions of the texts, except that they had different accent distributions according to the gold standard, or to distributions as observed in the productions of the two newsreaders. A perception study revealed that the versions of the bad speaker were rated as being significantly worse than the other versions. The two studies thus show that variation in accent distribution can indeed affect the way spoken texts are assessed in terms of their perceived quality. © 2012 Acoustical Society of America.
[<http://dx.doi.org/10.1121/1.4751539>]

PACS number(s): 43.70.Fq, 43.70.Mn, 43.71.Hw, 43.71.Sy [SSN]

Pages: 2616–2624

I. INTRODUCTION

Not all speakers are equally good. For instance, at a scientific conference where one can witness many different speakers, there typically tend to be presenters who are engaging, whereas others are boring, irrespective of whether they are native speakers. To some extent, the quality of their speech is determined by the extent to which they adequately pronounce words, produce grammatically correct sentences, and make appropriate lexical choices. But presumably, the difference in speaking style between good and bad speakers is also partly related to differences in the way they supplement their utterances with appropriate prosodic structures. Whereas some speakers may attract the attention of the audience by using a natural expressive style, others may come over as tiresome because they speak in a monotone, put too many or too few pauses in their speech, are hesitant, or mispronounce words, and so on. While all this may be intuitively clear, we still lack exact details on why some speakers are judged as performing better than others, even though it has been argued for centuries that speakers need to pay attention to their prosodic structures (as well as nonverbal features, such as facial expressions and gestures) to make sure that a message comes across successfully (Crowley and Hawhee, 2009). The goal of this article is to shed light on

differences in appreciation that stem from variability in the placement of pitch accents, which we analyze in spoken data of Flemish newsreaders as they cast the news on public radio. Flemish is the variant of Dutch that is spoken in the northern part of Belgium. It has a prosodic structure which is assumed to be the same as that of the variant spoken in the Netherlands (Collier and 't Hart 1981), and so far there are no studies which indicate otherwise.

Already in ancient Greece and Rome, it was recognized that a speaker's style may be as important as the actual content of his or her message in order to win over an audience. Quintilianus, in particular, remarked that nonverbal features, including prosodic characteristics such as intonation and rhythm, should be congruent with the content of the spoken message. By that he not only meant that a speaker's expressive style should match the emotional content of the messages (e.g., facial expressions should be sad when the message is negative), but it should also be in line with the salience of the information in the discourse. Inspired by the latter, the current article explores to what extent the quality of a particular speaking style is reflected in the way speakers distribute prosodic accents in a spoken text. According to standard models of intonation (e.g., 't Hart *et al.*, 1990), accents in Dutch are primarily realized by means of melodic variation, through abrupt changes in pitch on designated syllables (see Kochanski *et al.*, 2005). (Note that we use the word accent in the technical sense to refer to prosodic prominences, and not a speaker's way of speaking as in the

^{a)}Author to whom correspondence should be addressed. Electronic mail: m.g.j.swerts@uvt.nl

case of a foreign or regional accent.) While all spoken languages in the world are likely to have variation between prominent and less prominent words, there appear to be language-specific rules that dictate which words in a sentence should or should not be accented and how these accents relate to the information structure of the discourse (Ladd, 1996; Swerts *et al.*, 2002; Swerts, 2007). For Dutch, the language we will be studying in this article, influential work on accent prediction comes from Gussenhoven (1984, 1992), who proposes a three-step procedure to predict accents in a spoken text. Ignoring specific details of his model, the core of his approach consists of (i) the specification of focused constituents in a sentence, based on the salience of information in the discourse (e.g., related to given/new distribution), (ii) the assignment of accents on the basis of the argument-predicate structure within those focused constituents, and (iii) the deletion of accents within a phonological phrase according to some rhythmic principles.

Gussenhoven's approach to accents in Dutch has been a source of inspiration for other accounts as well (e.g., Dirksen, 1992; Marsi, 2001; Quené and Kager, 1989; see Hirschberg, 1992, for American English). But there are at least two issues that remain unexplored in these studies. First, these studies dealing with accent prediction are similar in the sense that they attempt to model an ideal speaker, or lead to implementations in text-to-speech systems in order to produce utterances with "perfect" accent distributions. However, when looking at speech databases with multiple speakers, it becomes immediately clear that speakers can vary considerably in the extent to which they produce accents, as it has, for instance, been reported that newsreaders within the Boston Radio News Corpus tend to accent every second word, which is an unusually high number (Ostendorf *et al.*, 1995). This fact has led to models that distinguish between accents that are obligatory (that all speakers "should" produce) vs other accents that are optional (e.g., Marsi, 2004). Also, speakers, just as they are known to produce speech errors and other disfluencies, are likely to produce "wrong" accent patterns as well. The models described above do allow for some variation. For instance, the rhythm rule of Gussenhoven (see also Marsi, 2001) is one which deletes accents that are surrounded by accents within the same phonological phrase, which can be applied a number of times. Since this rule is optional, it would be able to generate different versions of accent distributions for a similar text.

This brings us to a second issue, i.e., that it is still unclear how listeners evaluate texts with different accent distributions. While not many facts are known on how exactly prosodic structures affect the listeners' appreciation of spoken texts, there is circumstantial evidence that prosody matters to some extent. For example, there are studies in the area of speech synthesis that reveal that a prosodic module that correctly assigns accent structures is important for the synthesized text to be judged as natural and clear (Quené and Kager, 1989; Quené and Dirksen, 1990; Marsi, 2004), confirming earlier results with manipulated data that listeners prefer utterances with correct accent distributions (Nooteboom and Kruij, 1987). Ideally, a system should generate all the obligatory accents and can optionally insert others,

but may not produce accents that are incorrect. More generally, it has even been argued that listeners become more demanding about the goodness of the intonation module, when the segmental quality of the synthetic speech is of a good quality (Terken and Lemeer, 1988). These studies tend to have in common that they evaluate, through statistical comparison of accent distributions and perceptual assessment tests, how their predicted model compares to what a professional speaker would do, assuming that this speaker functions as a role model whose prosodic behavior could be mimicked. However, it is unknown how representative such a speaker is, and whether speakers whose accent patterns differ would be differently appreciated for that reason.

In addition, there is a tradition of psycholinguistic research on how correct or incorrect accent structures may affect a listener's processing of incoming speech. A typical finding, as, for instance, reported by Terken and Nooteboom (1987), is that listeners exhibit slower reaction times when they have to process utterances in which accents occur on given information rather than new information. More recent studies have found similar processing effects using eye-tracking techniques within the visual world paradigm, both with natural recordings (Dahan *et al.*, 2002; Ito *et al.*, 2011) and synthesized ones (van Hooijdonk *et al.*, 2007). Those processing results suggest that speakers who produce wrong accent distributions are more likely to be assessed as performing worse. Finally, from research on the L2 acquisition of prosody, it has been shown that prosodic—including accentual—errors have an important negative impact on native listeners' perceptual judgments of speech quality and intelligibility, as well as on their appreciation of the non-native speaker (Trouvain and Gut, 2007; Swerts and Zerbian, 2010).

The current article focuses on the speaking styles of Flemish newsreaders as they cast the news on Flemish radio. The choice to look at such speaker data is motivated by two facts. First, newsreaders have often served as role models within a specific language community (Swerts and Krahmer, 2010), and their speech data have been used as input for building speech synthesis systems. For instance, Philip Bloemendal, who was well-known as the person who presented the Polyglot News in Dutch movie theaters, has lent his voice to the Dutch speech community for making a diphone synthesis system. So, given that newsreaders tend to represent a role model for speech, it makes sense to explore how they produce their prosodic structures. But second, somewhat ironically, the public opinion often tends to differentiate between what people consider as good and bad speakers. This is, for instance, clear from yearly poll ratings organized by the website Radiovisie.eu (an independent Flemish-Dutch organization that reviews radio shows and programs) that show that people appreciate some newsreaders better than others.

The remainder of this article is organized as follows. First, we present a study with analyses of natural recordings, where we compare the prosodic structures in spoken news bulletins of two Flemish professional newsreaders, who were hypothesized (based on poll ratings) to represent a good vs bad speaker. The term "bad" is used as a shortcut to

refer to the speaker who received relatively low ratings in the yearly poll organized by the website *Radiovisie.eu*, but the term does not imply to mean that this speaker does not qualify as a newsreader. As a matter of fact, given that the procedure for selecting newsreaders on Flemish radio is quite strict, the speakers analyzed in the following studies should be considered excellent professional speakers, though there may be differences in general appreciation between them. Their speaking styles are first judged in a perception experiment, and subsequently their accent distributions are compared to see how well these match predictions of a gold standard provided by two independent experts. Second, we run an additional perception experiment which uses synthesized versions of the news bulletins of the two newsreaders. The synthesized versions of the news items are basically identical, except that they differ in accent distributions, with one version having accents according to the gold standard, and the other versions having accents according to what the good and bad speaker had produced. Both studies provide evidence that the quality of a person's speaking style is reflected in the way he or she distributes accents in a spoken text.

II. ANALYSES OF NATURAL RECORDINGS

A. Goal

The aim of this part of the study is to explore to what extent differences in the perceived quality of spoken news items (as assessed by a group of listeners) is related to prosodic characteristics of those spoken messages. In particular, we look at the extent to which the goodness of a spoken text can be explained by the extent to which the accent distribution in a text matches the one prescribed by a prosodic gold standard.

B. Stimuli

As a basis for our data, we collected speech materials from two different Flemish newsreaders. The two speakers (a female and a male speaker) were selected as they were rated quite differently by voters for a poll of the *Radiovisie.eu* that yearly gives different kinds of awards, including prizes for best newsreader. For our analyses, we decided to select a newsreader (male) whom the voters of *Radiovisie.eu* considered to be the best newsreader in the year 2008 (receiving 14.8% of the vote), and another newsreader (female) who did not appear in the top-10, and received less than 4% of the vote. The criteria that were used by the original voters are not explicitly stated, so that it remains to be seen to what extent the quality of the newsreaders is determined by prosodic features as well. In the remainder of this article, we will refer to these two newsreaders as the good and the bad speaker, respectively. For comparative reasons, we used the following procedure for data collection. First, we recorded a random sample of news items in June 2009 of the bad speaker who happened to be reading the 6:00 pm news on Flemish radio for a whole week. In addition, we contacted the good speaker to ask whether he would be willing to read those news items again. Note that we did not give

him very specific instructions, other than that he was asked to read the bulletins as he would do it on the radio. He was also not informed about the fact that we were going to use his analyses for comparative research. While the latter recordings were not aired in public, they were recorded in exactly the same studios of the Flemish public broadcasting company, so that they are very comparable in general audio quality. A typical news bulletin would consist of separate news items on different topics. A typical example of a news item is given in example (1), together with its translation into English (1').

- (1) De man die vanmorgen in Ledeberg zijn ex-partner doodschoot, had gisteren al een brutale overval gepleegd op twee vrouwen in Destelbergen. Hij probeerde vanmorgen na een mislukte ontsnapping zelfmoord te plegen en is nu in coma. Trui de Maré met de nieuwste informatie.
- (1') The man who fatally shot his ex-partner this morning in Ledeberg had already brutally robbed two women in Destelbergen the day before. He tried to commit suicide this morning after he failed to escape, and is now in a coma. Trui de Maré with the latest information.

The other news bulletins concerned a broad range of topics from daily reports of the Dauphiné Libéré cycling event to upcoming elections in Iran. For the analyses of the first study, we randomly selected 20 different news items.

C. Procedure

Forty-six listeners (invited through the social network of the School of Humanities at Tilburg University) were invited to participate in a web-based perception experiment on a voluntary basis. The listeners mainly included students from the humanities faculty, as well as friends and acquaintances of colleagues of the department, and were all "naïve" with respect to prosodic research. While most of the listeners were Dutch, some had a Flemish background. Inspection of the data did not reveal essential differences in assessment scores due to the participants' linguistic background. Twenty-two of the participants were presented with the audio files of 20 news items as spoken by the bad speaker, 24 heard the same 20 news items spoken by the good speaker. In other words, the experiment had a between-subject design, as we wanted to avoid that listeners would be able to compare the speakers themselves. Each participant received the news items in a different random order, to compensate for possible learning effects. We decided to present complete paragraphs [like the one in example (1) in Sec. IIB], rather than individual sentences, as the absence or presence of an accent has been argued to be related to the larger discourse structure of a text, so that the effect of different accent distributions may not be very clear when individual utterances would be presented (see also Terken and Collier, 1989). The listeners only heard the audio files, and did not have access to a written display of the text. The task given to listeners was to rate each spoken news item as to how well it was read. They were instructed to express their assessment on a 10-point scale, with 10 meaning "very

good” and 1 meaning “very bad,” and the numbers in between to represent values in between those two extremes. The experiment was self-paced, but listeners were encouraged not to reflect too much on their judgments and give an immediate response. On average, the experiment lasted approximately 10 min.

D. Results

The average ratings of the participants were analyzed by means of a repeated measures analysis of variance, with speaker as the between-subject factor (2 levels: good speaker, bad speaker) and the news item as within-subject factor (20 levels). There was a main effect of speaker [$F_{(1,44)}=5.856, p<0.05, \eta_p^2=0.117$] and of news items [$F_{(19,836)}=2.512, p<0.001, \eta_p^2=0.054$], whereas the interaction between speaker and news item turned out not to be significant. The speaker effect was such that the news items delivered by the bad speaker were considered to be slightly less good overall than those delivered by the good one [bad speaker: 6.577 (standard error: 0.201); good speaker: 7.252 (standard error: 0.193)]. Figure 1 visualizes the average scores for each separate news item for the two speakers, which shows that the ratings of the good speaker are always higher for each individual news item than the ratings of the bad speaker. Pair-wise comparisons using the Bonferroni-method revealed that the effect of news item was due to text number 19 receiving a significantly lower score than texts 1 and 20.

Using PRAAT, a software program to record, analyze, and manipulate speech, the recorded speech materials of both newsreaders were annotated by the first author in terms of the occurrence of clear pitch accents. That labeling was limited to a specification of presence or absence of pitch accents, and did not further specify the tonal configurations of these accents. Inspired by a method outlined in Collier and 't Hart (1981), the speech of the newsreaders was slowed down, as it highly facilitates the annotation procedure, because the beats can more easily be discerned. The procedure to annotate recordings in slow motion is also often used in facial coding procedures (see, e.g., Ekman and Rosenberg, 2005). In order to check whether the annotations of the accents in the spoken texts were reproducible, we randomly selected ten spoken paragraphs (five from each speaker) and gave them to an independent labeler, different from the two

authors. The independent labeler was also given an example paragraph with annotations to give an indication as to how accented words had been labeled in the data. When comparing the annotated subset of the speech data with that of the first author, there appeared to be substantial agreement between the two labelings (with a Kappa-score of 0.706).

The text version of the selected news items were also presented to two independent prosodic experts who were invited to annotate the texts in terms of presence or absence of accents without having access to the original recordings of the two newsreaders. In particular, they were asked to imagine how a professional newsreader or faultless speech synthesis system would have to produce accents in those utterances. We only did this for 19 texts because it turned out that the versions of one text produced by the good and bad speaker had one sentence with a slightly different wording and syntactic structure; thus, the accent distributions could not be compared. It should be noted that the gold standards produced by the prosodic experts were not based on any particular linguistic or psycholinguistic theory. The prescribed accents should therefore not be considered as predictions derived from a linguistic or computational model of accent placement. Accordingly, our experiments are not intended to directly evaluate any such model. Instead, the gold standard reflect the experts’ knowledge on the best use of prosody as accumulated through observing and analyzing the prosody of speakers on a daily basis. Admittedly, this includes broadcaster’s speech, both good and bad, but it is based on a much wider range of observations.

For the remainder of this article, we will consider these two annotations as gold standards for accent distribution. We took the annotations of one of the experts as a reference (Gold standard 1) and used the other one (Gold standard 2) as a way to check whether the accentual specifications were reproducible.

Table I gives confusion matrices that compare different accent distributions as predicted by one of our experts [Gold standard 1 (GS1)] with those of the other expert, and with the distributions produced by the two selected speakers, and also gives the respective kappa values that express the agreement between these different distributions. As can be seen, the kappa score for the comparison between the two gold standards is the highest of all, suggesting that the experts have a very clear view on what is deemed good accentuation.

FIG. 1. Average assessment scores for 20 texts produced by two different speakers.

TABLE I. Comparison of presence and absence of accents according to Gold standard 1 with distributions of Gold standard 2, and in the recordings of the good speaker and the bad speaker (actual countings and row percentages), together with Kappa values.

Model	Accent according to Gold standard 1			Kappa
	Accent	Yes	No	
Gold standard 2	Yes	530 (86.3%)	84 (13.7%)	0.737
	No	32 (9.8%)	294 (90.2%)	
Good speaker	Yes	535 (87.1%)	79 (12.9%)	0.642
	No	74 (22.7%)	252 (77.3%)	
Bad speaker	Yes	530 (86.3%)	84 (13.7%)	0.571
	No	97 (29.8%)	229 (70.2%)	

When comparing the accent distributions of Gold standard 2 (GS2) with those of the newsreaders, we get basically near-identical results, with a relatively low kappa score of 0.615 for comparisons between the bad newsreader and that standard, and of 0.722 for comparisons with the good speaker. Closer inspection of the respective annotations suggested that the two experts agreed almost always on the occurrence of nuclear accents, but differed regarding prenuclear ones, which different models would indeed consider to be optional. In addition, the kappa values for the agreement between GS1 and the distributions of the two newsreaders appear to be consistent with the difference in perceived quality of their speaking style; the kappa value for the bad speaker, while still referring to a fair agreement, turns out to be lower than that for the good speaker. To further support this, we performed a repeated measures analysis of variance (ANOVA) where we analyzed the kappa values for the 19 texts that compared the distributions of GS1 with that of GS2, the good speaker and the bad speaker. This analysis revealed that the differences between the kappa values for these three comparisons were statistically significant [$F_{(2,17)} = 17.813$, $p < 0.001$, $\eta_p^2 = 0.677$]. Pairwise comparisons using the Bonferroni method revealed that each of the pairs were significantly different from each other, meaning that the kappa values of the bad speaker were significantly lower than those for the good speaker, while both were lower than those for the comparison between the two gold standards.

E. Discussion

The analyses of natural recordings revealed that speakers who were independently rated as being good or bad also

turned out to produce different accent distributions. In particular, the accents of the good speaker were a better match to what a gold standard (coming from two independent intonation experts) had predicted. Closer analysis of the data suggests that listeners gave especially bad scores to those texts that exhibit cases of unaccented words that appeared in nuclear position.

To give some illustrative examples, Table II shows a case of a news item with accents (words that are underlined) in the speech of the bad speaker, the good speaker, and as predicted by the Gold standard 1. We also added an English translation of the Dutch text. Looking closely at the example suggests that the inadequate accent distribution in the speech of the bad speaker could be partly responsible for a low score in appreciation of this text. In particular, the distributions in the text produced by the good speaker and in the version prescribed by the gold standard agree regarding the presence of a nuclear accent on “presidentsverkiezingen,” on “Ahmadinejad,” and on “Mousavi,” which are absent in the bad speaker version, which renders a rather marked rhythmic structure on this. As indicated in the Introduction, variation between accent distributions is not necessarily bad, as even the two experts producing the gold standards can generate (slightly) different accent distributions. However, those expert predictions are more in line with the rhythm rule of Gussenhoven in the sense that it allows multiple distributions because of the fact that the accent deletion procedure is optional and recursive. Yet, the variable distributions obtained this way have to do with presence or absence of prenuclear accents, whereas the gold standards tend to agree mostly on the occurrence of nuclear accents. If a speaker leaves out such a nuclear accent (as in the example of Table II), that is a serious violation of the rhythm rule, and leads to a pattern which normally is reserved for contrastive accent interpretations: if a nuclear accent is moved away from its default final position within a phrase, it often leads to a contrastive interpretation (see Ladd, 1996; Krahmer and Swerts, 2001). And such a contrastive reading of the text is infelicitous in the text at hand, which probably has led to a rather negative score for the bad speaker.

Obviously, while the analyses in terms of the kappa-scores suggest that the differences in accent distribution may be responsible for the differences in ratings, additional work is needed to find out to what extent this is a sensible claim. Indeed, the speech produced by both newsreaders may differ in other characteristics as well, including the way specific

TABLE II. Examples of original Dutch utterances, and the English translation, with accent distributions (accented words are underlined) according to the bad speaker, the good speaker, and the Goldstandard 1. Further explanations in the text.

Original	Bad speaker	Over zowat een half uur zouden <u>in Iran</u> de <u>stembussen</u> dichtgaan voor de presidentsverkiezingen. Er zijn <u>vier kandidaten</u> , maar de race <u>gaat</u> vooral tussen de <u>huidige</u> president Ahmadinejad en de <u>hervormer</u> Mousavi. De <u>opkomst</u> is <u>overweldigend</u> .
	Good speaker	Over zowat een <u>half uur</u> zouden in Iran de <u>stembussen</u> dichtgaan voor de <u>presidentsverkiezingen</u> . Er zijn <u>vier kandidaten</u> , maar de <u>race</u> gaat vooral tussen de <u>huidige</u> president <u>Ahmadinejad</u> en de <u>hervormer</u> Mousavi. De <u>opkomst</u> is <u>overweldigend</u> .
	Gold standard 1	Over zowat een <u>half uur</u> zouden in <u>Iran</u> de <u>stembussen</u> dichtgaan voor de <u>presidentsverkiezingen</u> . Er zijn <u>vier kandidaten</u> , maar de <u>race</u> gaat <u>vooral</u> tussen de <u>huidige</u> president <u>Ahmadinejad</u> en de <u>hervormer</u> Mousavi. De <u>opkomst</u> is <u>overweldigend</u> .
Translation		In approximately half an hour the voting booths in Iran for the presidential elections will be closed. There are four candidates, but the contest runs primarily between the current president Ahmadinejad and the reformer Mousavi. The attendance is overwhelming.

words are pronounced. Also, the newsreaders differ in gender, which possibly may have affected the judgments of some observers as well, e.g., in case they have a preference for either a male or a female voice. Specifically with respect to prosody, there may have been differences in intonational phrase boundaries, which in turn affect accent placement. For example, speakers try to avoid intonational phrases that do not contain at least one accent. The presence of an intonational boundary may therefore necessitate additional accents. Therefore, in order to further investigate to what extent accentual differences can be held responsible for the variation in the way the speaking styles are assessed, we conducted another perception experiment in which synthetic stimuli were tested. The accentual properties of the synthetic stimuli were systematically manipulated according to the different speaker models. This allowed us to control for prosodic factors such as intonational phrasing and speech rate.

III. PERCEPTION OF SYNTHESIZED STIMULI

A. Goal

The aim of the second part of this study was to see whether additional evidence could be found for the claim that variation in accent distribution can be held responsible for perceived differences in the quality of the spoken news items. Therefore, we conducted an additional perception experiment in which listeners were presented with three different versions of the same text that varied in accent distribution, i.e., according to a good speaker model, a bad speaker model, and a gold standard model.

B. Stimuli

The stimuli for the second experiment were made with Nextens, a freely available open source text-to-speech system for Dutch based on the Festival TTS platform (Taylor *et al.*, 1998). As input we used the news text together with an aligned specification of the intonation according to the ToDI system (Gussenhoven, 2005), which is an autosegmental description of Dutch intonation along the lines of the ToBI system for transcribing American-English intonation. Briefly explained, ToDI allows symbolic transcription of

pitch accents and intonational boundaries in terms of combinations of *H* and *L* tones, representing high and low pitch targets, respectively. For example, *H*L* is a falling pitch accent where the star suffix indicates that the *H* tone is timed on the accented syllable. Similarly, *%L* and *H%* represent a low start and high end of an intonational phrase where the percent affix indicates that they are associated with an intonational phrase boundary. The shape of an utterance's complete intonational tune is determined by the consecutive contour shapes defined by its pitch accents and boundary tones. The intonational tune was essentially the same for all texts, except that the distribution of accented words was modeled on the basis of the productions of the good and bad speaker, and according to our Gold standard 1. The pitch contour follows the most commonly used, neutral pattern in Dutch: phrases always start with an initial low boundary tone (*%L*), the first accent in the phrase is realized as a "pointed hat" (*H*L*), subsequent accents in the same phrase accents become a downstepped pointed hat (*!H*L*), sentence-internal phrases end with a continuation rise (*H%*), and sentence-final phrases end with a final low boundary tone (*L%*). The abstract intonational description was translated to an *F*₀ contour using a target interpolation model for Dutch intonation (van den Berg *et al.*, 1992). The generated contours thus appear to resemble those in the study by Terken and Collier (1989). Figure 2 gives an example of a sentence with a generated pitch contour. Intonational phrasing was based on punctuation (i.e., every comma becomes a phrase boundary) as well as length considerations (i.e., very long phrases were avoided by inserting a break at an appropriate location). This phrasing was identical across all three synthesized versions. Likewise durational structure and speech rate were constant across all versions, except for very minor lengthening of accented syllables. Errors in grapheme-to-phoneme conversion were manually corrected by adding new entries to the lexicon. The final waveform synthesis relied on the MBROLA diphone synthesis system, using the female NL3 voice (Dutoit *et al.*, 1996). Even though diphone synthesis may not deliver the most natural sounding speech in comparison with unit synthesis, it guarantees that, except for their intonation, all three versions

FIG. 2. Waveform and corresponding *F*₀-contour of the Dutch sentence “De twee raakten zwaargewond, maar zijn buiten levensgevaar (The two got seriously wounded, but do not have to fear for their lives)” with ToDI labels and IPA (International Phonetic Alphabet) transcription (further explanations in the text).

are identical (apart from the marginal durational differences mentioned), and the speech quality remains constant throughout the utterance.

Given that we wanted to apply a within-subject design, we synthesized only a subset of ten texts that were presented in the first study in order to reduce experimental time for listeners. As we were particularly interested in the effects of differences in accent distribution, we chose to produce versions of the text according to Gold standard 1, and further selected texts where the difference in accent distributions of the two newsreaders was the largest in terms of the kappa statistics reported above. This would then produce 30 texts in total.

C. Procedure

Forty-one participants (again from the social network of Tilburg University) took part in a perception experiment on a voluntary basis. None of the participants had taken part in the first perception experiment with natural stimuli. They were presented with all the 30 synthesized news items described in Sec. III B. Their task was to assess the perceived quality of the spoken paragraphs by giving a score on a 10-point scale, where “1” represents a bad quality and “10” represents a very good quality. To compensate for possible order effects, each participant got a differently randomized list of the speech stimuli. Unlike the procedure of the perception experiment with natural recordings in the first study, the current experiment had a complete within-subject design.

D. Results

The data were analyzed with a repeated measures ANOVA with speaker model (three levels: gold standard, good speaker, bad speaker) and text (ten levels) as independent within-subject factors, and the scores on the 10-point scale as dependent factor. The analysis revealed a significant main effect of speaker [$F_{(2,80)} = 19.425, p < 0.001, \eta_p^2 = 0.327$] and text [$F_{(9,360)} = 13.818, p < 0.001, \eta_p^2 = 0.257$], while the interaction between text and speaker was also significant, though with a very low effect size [$F_{(18,720)} = 6.096, p < 0.001, \eta_p^2 = 0.063$]. Table III gives the mean scores for the different intonation models. Pairwise comparisons using the Bonferroni-method indicated that the scores for the bad speaker model were significantly lower than the scores for both other models, which were not significantly different. The effect of text was due to the fact that some texts overall scored worse or better than others.

If we rank the average scores for each text in terms of how the models of the bad speaker, the good speaker, and the gold standard performed, we get the distribution of

TABLE III. Listeners’ assessment scores for different prosodically manipulated versions of news items (synthesized voices).

Model	Score (standard error)	<i>F</i> -stats
Bad speaker	4.715 (0.193)	$F_{(2,80)} = 19.425, p < 0.001, \eta_p^2 = 0.327$
Good speaker	5.261 (0.175)	
Gold standard	5.161 (0.180)	

scores as given in Table IV. As can be seen, for seven out of ten texts, the bad speaker model performs worst of all three models, and never gets the highest score. Conversely, the models of the good speaker and the gold standard are rarely assessed as being the worst, and score as first or second best model.

E. Discussion

The outcome of the experiment with synthetic stimuli confirms the claims based on the findings of the first study that appropriate or inappropriate accent distributions can have a positive or negative effect on how a particular spoken message is appreciated by a listener. Interestingly, the good speaker and gold standard models get scores that are very comparable, and together outperform the bad speaker model. As the different versions of the texts were controlled for type of intonation pattern, speech tempo, and voice quality, it is remarkable to see that a mere change in accent distribution can have significant repercussions for the perceived quality. Note that the overall assessment scores for the speaker models are comparatively lower than the ones for the natural data discussed in the first experiment, which is probably due to the fact that the segmental quality of the speech synthesis is of a lower quality than that of the natural speech stimuli used in the first experiment.

IV. GENERAL DISCUSSION

The current article has reported evidence that the way accents are distributed in a spoken text affects how well such a text is appreciated by listeners. We have shown this through analyses of natural recordings of two Flemish newsreaders as they cast the news on public radio, which revealed that the perceived quality of the spoken texts was reflected in the extent to which the accents in the texts matched a gold standard, as provided by independent intonation experts. A follow-up experiment with synthesized stimuli supported the earlier finding as it turned out that synthesized texts with accent distributions of the gold standard or of a good speaker received significantly better ratings from listeners than texts with accent distributions of a bad speaker. Note that despite the somewhat dismissive use of the term “bad,” it is interesting to note that even the speaker who performed more poorly is a professional newsreader, so that the problem could be even worse with speakers who have less experience in reading text.

Obviously, the research could be extended in a number of ways. This study only looked at the effects of the correct or incorrect placement of prosodic accents on listeners’

TABLE IV. Rank order of the scores for the models of the good speaker, the bad speaker, and the gold standard in terms of whether these scored best, worst, or in the middle.

Model	Worst score	Middle score	Best score
Bad speaker	7	3	0
Good speaker	2	3	5
Gold standard	1	4	5

appreciation of spoken text. While we did find that differences in distribution correlate with the perceived goodness of a text, there are more prosodic dimensions that are potentially relevant for an appropriate speaking style. One could think of variation in melodic patterns, rhythmical and temporal structures, and differences in the placement of pauses. We expect, however, that the newsreader data we analyzed for this study are not very suitable for exploring differences in perceived naturalness due to phrasing differences because, unlike the variation in accent distribution, the placement of major prosodic boundaries is likely to be quite comparable for the two speakers due to the fact that these are highly driven by the punctuation of the text. In addition, an interesting alley to pursue further is to try and also “repair” the speech produced by a bad speaker. In order to do so, one can imagine using modern techniques of prosodic transplantation (copying intonation patterns from one utterance on another utterance), especially in combination with up-to-date waveform manipulation software that guarantees that the manipulated speech materials preserve their naturalness. This could take at least two forms. One approach would be to change accent distributions in a spoken text according to what various prosodic models would predict, in order to find out whether that indeed ameliorates the quality of the speaking style, and if so, it would be relevant to know which of the models scores best in that respect. Another approach would be to copy the accent distributions of a good speaker onto the speech of the bad speaker, and vice versa, to see whether that leads to improvement or degradation in the perceived quality of the speech.

And finally, the results of the two studies presented above relate to “naïve” listeners’ naturalness ratings, which are of course rather metalinguistic in nature, in the sense that people are asked to reflect on the language of a speaker. It could therefore also be useful to supplement these studies with more functional experiments that test to what extent the actual processing of the speech is positively or negatively affected by good or bad accent distributions. It is of course known from psycholinguistic studies with controlled utterance stimuli that participants’ processing time and eyegaze patterns can be influenced by presence or absence of accents (Cutler *et al.*, 1997; Terken and Nootboom, 1987; Dahan *et al.*, 2002; Ito *et al.*, 2011; van Hooijdonk *et al.*, 2007). It would be interesting to see whether such findings generalize to more natural data (e.g., from newsreaders) to see whether people have a harder time understanding news items produced by a “bad” speaker, or remembering specific details of the content of these texts. Another extension would be how good or bad accentuation affects the quality of speech produced by L2 speakers.

ACKNOWLEDGMENTS

We thank Pauline Heinrichs, Marieke Hoetjes, Kitty Leuvenink, Madelène Munnik, Jacqueline Dake, and Lennard van de Laar for their help with transcriptions, and for help with setting up and conducting the listener experiments. Carlos Gussenhoven and Jacques Terken are acknowledged for being willing to provide the two prosodic gold standards.

Parts of this research have previously been presented at the “International Workshop on Tone and Intonation” in September 2011 in Nijmegen, The Netherlands.

- Collier, R., and 't Hart, J. (1981). *Cursus Nederlands Intonatie (Course on Dutch Intonation)* (Acco, Louvain), pp. 1–88.
- Crowley, S. and Hawhee, D. (2009). *Ancient Rhetorics for Contemporary Students* (Pearson, New York), pp. 1–462.
- Cutler, A., Dahan, D., and van Donselaar, W. (1997). “Prosody in the comprehension of spoken language: A literature review,” *Lang. Speech* **40**(2), 141–201.
- Dahan, D., Tanenhaus, M. K., and Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *J. Mem. Lang.* **47**(2), 292–314.
- Dirksen, A. (1992). “Accenting and deaccenting: A declarative approach,” in *Proceedings of the 14th Conference on Computational Linguistics (COLING'92)*, Nantes (France), August 23–28, pp. 865–869.
- Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and Van der Vrecken, O. (1996). “The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes,” in *Proceedings of the Fourth International Conference on Spoken Language (ICSLP 96)*, pp. 1393–1396.
- Ekman, P. and Rosenberg, E. (Eds.) (2005). *What the Face Reveals*, 2nd ed. (Oxford University Press, New York), pp. 1–639.
- Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents* (Foris, Dordrecht), pp. 1–352.
- Gussenhoven, C. (1992). “Sentence accents and argument structure,” in *The-matic Structure. Its Role in Grammar*, edited by I. M. Roca (Foris, Berlin), pp. 79–106.
- Gussenhoven, C. (2005). “Transcription of Dutch intonation,” in *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, (Oxford University Press, Oxford), pp. 118–45.
- Hirschberg, J. (1992). “Using discourse context to guide pitch accent decisions in synthetic speech,” in *Talking Machines: Theories, Models and Designs*, edited by G. Bailly, C. Benoit, and T. R. Sawallis (Elsevier, Amsterdam), pp. 181–184.
- Ito, K., Jincho, N., Minai, U., Yamane, N., and Mazuka, R. (2011). “Intonation facilitates contrast resolution: Evidence from Japanese adults and 6-year olds,” *J. Mem. Lang.* **66**(1), 265–284.
- Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). “Loudness predicts prominence: fundamental frequency lends little,” *J. Acoust. Soc. Am.* **118**, 1038–1054.
- Krahmer, E. J. and Swerts, M. (2001). “On the alleged existence of contrastive accents,” *Speech Commun.* **34**(4), 391–405.
- Ladd, D. R. (1996) *Intonational Phonology* (Cambridge University Press, Cambridge), pp. 1–349.
- Marsi, E. (2001). *Intonation in Spoken Language Generation* (LOT, Utrecht), pp. 1–347.
- Marsi, E. (2004). “Optionality in evaluating prosody prediction,” in *Proceedings of 5th ISCA Speech Synthesis Research Workshop*, Pittsburgh, pp. 13–18.
- Nootboom, S. G., and Kruij, J. G. (1987). “Accents, focus distribution, and the perceived distribution of given and new information: An experiment,” *J. Acoust. Soc. Am.* **82**, 1512–1524.
- Ostendorf, M., Price, P. and Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus, Technical Report Number ECS-95-001 (Boston University Press, Boston, MA), pp. 1–19.
- Quené, H., and Dirksen, A. (1990). “A comparison of natural, theoretical and automatically derived accentuations of Dutch texts,” in *Proceedings ESCA Workshop on Speech Synthesis*, September 25–28, Autrans, France, pp. 137–140.
- Quené, H., and Kager, R. (1989). “Automatic accentuation and prosodic phrasing for Dutch text-to-speech conversion,” in *Proceedings European Conference on Speech Communication and Technology (Eurospeech 1989)*, Edinburgh, Scotland, pp. 214–217.
- Swerts, M. (2007). “Contrast and accent in Dutch and Romanian,” *J. Phonetics* **35**(3), 380–397.
- Swerts, M., and Krahmer, E. J. (2010). “Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience,” *J. Phonetics* **38**, 197–206.
- Swerts, M., Krahmer, E. J., and Avesani, C. (2002). “Prosodic marking of information status in Dutch and Italian: A comparative analysis,” *J. Phonetics* **30**(4), 629–654.

- Swerts, M., and Zerbian, S. (2010). "Intonational differences between L1 and L2 English in South Africa," *Phonetica* **67**, 127–146.
- Taylor, P., Black, A., and Caley, R. (1998). "The architecture of the Festival speech synthesis system," in *The Third ESCA Workshop in Speech Synthesis*, pp. 147–151.
- Terken, J., and Collier, R. (1989). "Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch," in *Proceedings Eurospeech. First European Conference on Speech Communication and Technology*, September 27–29, Paris, France, pp. 1357–1359.
- Terken, J., and Lemeer, G. (1988). "Effects of segmental quality and intonation on quality judgments for texts and utterances," *J. Phonetics* **16**(4), 453–457.
- Terken, J., and Nootboom, S. G. (1987). "Opposite effects of verification latencies of accentuation and deaccentuation for given and new information," *Lang. Cognit. Processes* **2**(3/4), 145–163.
- 't Hart, J., Collier, R., and Cohen, A. (1990). *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody* (Cambridge University Press, Cambridge), pp. 1–212.
- Trouvain, J. and Gut, U. (2007). "Non-native prosody: Phonetic description and teaching practice," *Trends in Linguistics: Studies and Monographs (TiLSM)* No. 186 Mouton de Gruyter, Berlin).
- van den Berg, R., Gussenhoven, C. and Rietveld, T. (1992). "Downstep in Dutch: Implications for a model," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (Cambridge University Press, Cambridge), pp. 335–359.
- van Hooijdonk, C. M. J., Commandeur, E., Cozijn, R., Krahmer, E., and Marsi, E. (2007). "Using eye movements for online evaluation of speech synthesis using," in *Proceedings Interspeech 2007*, August 27–31, Antwerp, Belgium, pp. 1346–1349.