Modelling personality features by changing prosody in synthetic speech

Jürgen Trouvain[1,2], Sarah Schmidt[3], Marc Schröder[4], Michael Schmitz[3] & William J. Barry[2]*

[1]Phonetik-Büro Trouvain, Saarbrücken
[2]Institute of Phonetics, Saarland University
[3]Institute of Computer Science, Saarland University
[4]DFKI GmbH, Saarbrücken
trouvain@phonetik-buero.de| sarah@xantippe.cs.uni-sb.de |schroed@dfki.de|
schmitz@cs.uni-sb.de|wbarry@coli.uni-sb.de

Abstract

This study explores how features of brand personalities can be modelled with the prosodic parameters pitch level, pitch range, articulation rate and loudness. Experiments with parametrical diphone synthesis showed that listeners rated the prosodically changed versions better than a baseline version for the dimensions "sincerity", "competence", "sophistication", "excitement" and "ruggedness". The contribution of prosodic features such as lower pitch and an enlarged pitch range are analyzed and discussed.

1. Introduction

In several studies it has been found that prosodic characteristics are linked with features of personality. E.g. the habitual speech rate of an individual speaker can have an influence on the benevolence perceived by listeners [2,21]; excitement or arousal is strongly correlated to prosodic features such as pitch level [23], pitch range [2,8] and tempo [2,7].

In synthetic speech, prosodic correlates of *emotions* have been modelled in a number of different ways (e.g., [4,6,18]). Aspects of both, personality and emotions can be subsumed under the heading "expressive speech synthesis", a term which also includes the expression of affect and attitude. In contrast to the modelling of emotions for computer-generated voices, little attention has been paid to the modelling of personality. Similarly, studies focussing on the prosody of personality features in human speech are much less frequent than those exploring the emotional side of human speech.

1.1. The "big five" dimensions of personality

In psychology the five factor model [9] is widely used to categorize the dimensions of personality. The "big five" are summarized in Table 1.

Studies investigating natural speech do not necessarily use these terms verbatim, but often investigate related concepts under a different name. For example, the dimension "agreeableness" is strongly related to the concept of politeness (e.g. [12]); the "energetic" end of the "extraversion" dimension is linked to the concept of high arousal; and the "careless" end of the "conscientiousness" dimension of personality can be assumed to correspond to a certain extent to an informal and sloppy speaking style, characterized by a high level of reductions, assimilations and elisions of sounds.

Table 1: *Dimensions of personality according to the five-factor model with example features for each dimension.*

Personality dimension	High level	Low level
Neuroticism	sensitive, nervous	secure, confident
Extraversion	outgoing, energetic	shy, withdrawn
Openness to experience	inventive, curious	cautious, conservative
Agreeableness	friendly, compassionate	competitive, outspoken
Conscientiousness	efficient, organized	easy-going, careless

Scherer [16] investigated vocal correlates of personality in a study of speech recorded from simulated jury discussions among male American speakers. When correlating peer-rated speaker personality with phoneticians' ratings of voice quality, he found significant positive correlations of the personality dimension of "emotional stability" (the inverse of "neuroticism") with vocal effort and dynamic range, and of "extraversion" with vocal effort, dynamic range and nasality. In other words, "American speakers rated as emotionally stable and extroverted by their peers seem to speak with a louder and possibly more nasal voice, using a greater range of loudness variation" [16: 473].

In one of the few studies that explicitly investigated correlates of personality in synthetic speech, Nass & Lee [10] tested the possibility of modelling the dimension of introversion~extroversion in synthetic speech by modifying the three prosodic parameters pitch range, pitch level, intensity and tempo. Their results showed that listeners perceived the resulting synthetic voices with the same degree of introversion as predicted by the model.

1.2. Marketing aspects of personality

Parallel to the dimensions of personality in humans there are also models which describe the personality of products and service brands. In the context of marketing research, Aaker [1] analyzed brand personality using a factorial analysis, and identified five dimensions of brand personality and their most typical attributes (Table 2).

* Preparation of this paper was supported by the EU Project HUMAINE (IST-507422).

Table 2: *Dimensions of brand personalities with typical attributes according to [1].*

Personality dimension	Attributes
Sincerity	down-to-earth, honest, wholesome, cheerful
Excitement	daring, spirited, imaginative, up-to-date
Competence	reliable, intelligent, successful
Sophistication	upper class, charming
Ruggedness	outdoorsy, tough

For those who are not experts in marketing, the meaning of these dimensions may become clear if they think of voices in advertisements on the radio and television. A low male voice is not suitable for a children's toy; conversely, a high female voice is not the voice to advertise a sporting utility vehicle. There is evidence that in radio ads a voice that fits the product helps people to remember the brand, the product and claims for that product [11].

1.3. Review of literature

But what does a voice sound like that expresses the qualities of a brand that is supposed to stand for "competence", i.e. which claims to be "reliable, intelligent, successful"? Very few voice studies deal with exactly the attributes used here to describe brand personality dimensions. We must therefore consider research that has aimed to investigate merely related attributes.

Sincerity can be linked to benevolence: high-pitched voices sound less benevolent, and an habitually slow speaking rate sounds cold [2,21]. Speech with a more variable intonation was evaluated as more benevolent than speech with less variable intonation [3]. Louder voices are judged to be friendlier than softer ones [15]. Deception is correlated with a rising F0 mean [22].

Excitement or arousal is correlated to a number of prosodic variables. The perception of high arousal is usually linked to high pitch level and range (e.g., [8,18], a louder voice with more high-frequency energy (e.g., [8,23]. In addition, it is often reported to be linked to fast speech rate [7], although the opposite was also reported [8].

Competence can be transmitted by speech rate. Faster speakers appear more convincing, more confident, more intelligent and more objective [2,21]. Less pauses and repeats and a more dynamic voice give the impression of competence [25]. High-pitched voices are judged as less competent whereas low pitch raises the degree of apparent confidence [2]. Furthermore, male voices are seen as more competent in connection with male subject matter such as technical and computer-related topics. Female voices are judged as more competent on topics such as personal relationships and the family [14]. [15] found that louder voices and male voices were considered more logical than softer voices and female voices.

With respect to sophistication the facets soft and attractive are best expressed with a voice which is identified as female or which sounds female-like. Voices are evaluated as attractive when they are less monotonous and very clear with a low pitch, and also when they have a larger pitch range with not too many and not too few pauses [25]. An upper-class way of speaking was regarded as more precise [3].

"Ruggedness" can be equated with sounding "robust", which can be achieved with a loud, low male voice speaking at a slow tempo. This dimension is linked in an obvious way with the so-called frequency code [13], which describes a universal relationship between the size of a speaker/animal and the fundamental frequency of its voice. Besides size, weight and robustness of the product correlate with these features.

Excitation is very similar to the concept of arousal, which has been modelled and tested in a synthesis context [18]. Higher arousal was realised as higher F0 level and range, more prominent accents, steeper F0 rises and falls, faster speech rate, more but shorter pauses, a longer duration of obstruent consonants compared to vowel durations, and a voice quality expressing high vocal effort [19]. These features were rated, in a listening test, as more appropriate for texts expressing the intended (high or low) arousal than for texts expressing the opposite degree of arousal.

2. Experiment

We carried out an experiment addressing the question whether it is possible to model the five different dimensions with the same voice, so that listeners judge them to be recognized as the intended personality dimension. A second question was whether listeners considered the modelled utterances as more suitable than a neutral baseline for the respective personality dimension.

2.1. Models for synthetic Speech

For the modelling of the personality dimensions we generated different versions of the test utterance by changing the four prosodic parameters pitch range, pitch level, tempo and intensity (cf . [10]). The German speech synthesizer Mary [20] using the MBROLA [5] voices de6 (male) and de7 (female) was used to generate the different versions of the test utterance by changing of:

- pitch range (in semitones),
- pitch level (in Hz),
- articulation rate (as a durational factor in ms),
- loudness as a loud, a soft or a modal "voice quality" (see [19] for more details).

The parameters were set to three levels: either it was unchanged, i.e. *on the same level* as the neutral baseline, or it was *higher* than the normal setting for the speech synthesizer, or *lower* than the normal setting (cf. Table 3). As the default tempo setting of the synthesizer used was rather slow, the rate of the baseline version was increased by 15%.

Table 3: *Three levels of manipulation of the prosodic parameters.*

manipulation level		pitch level	pitch range	tempo	loud-ness
lowered	-1	-30%	2 st	0%	soft
baseline	0	0%	4 st	+15%	modal
raised	+1	+30%	8 st	+30%	loud

The selection of the levels for each model was based on values from the literature (see Section 1.3). Table 4 lists the settings for the different models of the brand personality dimensions.

Table 4: *Dimensions of brand personality with the schematized levels of manipulation.*

Personality dimension	pitch level	pitch range	tempo	loudness
baseline	0	0	0	0
sincere	0	+1	0	+1
excited	+1	+1	+1	+1
competent	-1	+1	+1	+1
sophisticated	-1	+1	0	0
rugged	-1	0	-1	+1

2.2. Method

The task of the test that we report here was to judge how much the modelled utterances fit the five personality dimensions on a Likert scale ranging from 1 (does not fit at all) to 5 (fits very well). For each dimension prototypical attributes were offered.

36 native speakers of German (10 females, 26 males; 25 subjects 17-27 yrs, 11 subjects older than 23 yrs) completed an online test.

The test utterance contained four phrases with a total of 31 syllables: "Hallo, ich bin Produkt XY. Ich möchte mich kurz vorstellen. Ich werde nun meine Eigenschaften erläutern." ("Hello, I am product XY. I would like, briefly, to introduce myself. I shall now explain my features."). More details about this experiment can be found in [17]. Twelve audio files (5+1 dimensions x 2 voices) were generated with the Mary XML tool [20] for the male and the female voice.

2.3. Results

The results in Tables 5 and 6 show how well the modelled speech fits the respective personality dimension. It can be seen that in nearly all cases, the modifications made increased the ratings of the baseline voice with respect to the intended personality dimension (shaded cells).

Table 5: *Results for the male voice.*

assigned to	sincere	excited	compet.	sophist.	rugged
modelled as					
baseline	3.2	2.5	3.5	2.7	3.0
sincere	*3.5*	3.3	3.8	3.3	3.3
excited	3.3	***3.4***	3.5	2.8	2.8
competent	3.5	2.9	*4.0*	3.2	**4.1**
sophist.	**3.7**	2.3	**4.1**	***3.6***	3.6
rugged	2.6	1.5	3.4	2.6	*3.6*

Table 6: *Results for the female voice.*

assigned to	sincere	excited	compet.	sophist.	rugged
modelled as					
baseline	3.4	2.9	3.5	3.1	2.6
sincere	***3.5***	3.2	**3.6**	**3.4**	2.4
excited	2.8	***4.1***	2.7	3.2	2.2
competent	2.9	3.2	*3.3*	2.6	3.0
sophist.	3.2	2.5	3.5	*3.2*	**3.3**
rugged	2.6	1.8	2.8	2.2	*2.7*

This effect is particularly clear for the voice de6 (male), where the ratings of all dimensions improve by 0.3 to 0.9 between the baseline and the intended personality-specific prosodic configuration. For the voice de7 (female), however, only one dimension gains considerably from the rating of the intended category: the version intended to be excited is rated 1.2 points higher than the baseline. For three dimensions, the personality-specific versions are rated only marginally better than the baseline (+0.1), and the "competent" version is actually judged as less competent than the baseline (-0.2).

The best rated versions for the specific dimensions (in bold) were not always those of the intended models, e.g. for "rugged" the intended model scored behind the "competent"-model. The model for "sophisticated" scored best for the male voice in three dimensions, namely for "sophisticated" (as expected), for "sincere" and for "competent". Interestingly, for the female voice, the "sincere"-model is judged best for the same three dimensions and not only for the intended one.

Comparing the results for the baseline version of the two voices, the female voice was judged better than the male voice on three out of the five dimensions. However, a comparison of the best scores for each dimension shows that the male voice is judged better than the female voice in three dimensions, once equally as good ("sincere") and once the female voice is better than the male one ("excited"). This "excited" version is also one of the best three overall ratings together with competent and rugged for the male voice.

2.4. Discussion

This experiment demonstrates that for certain brand personality dimensions there are clear preferences for prosody modelled synthesis.

Comparing the scores of the baseline versions, the female voice scores better than the male voice except for "rugged", which was expected to have a prototypical male preference. However, the preferred model for "rugged" was the one for "competent" followed by the models for "rugged" and "sophisticated". The common prosodic feature of all three models is low pitch, while other parameters such as tempo are quite different for the three models.

"Excited" happens to be the lowest of the "best values" for the male voice (score: 3.4), but the highest "best value" of the female voice (score: 4.1). This finding for the male voice is slightly surprising: in [18] the evaluation by listeners was rather successful for the male voice "de6". In an additional, informal listening test, we could replicate the rather bad performance of the "excited"-model. It might be that changes in voice quality and those in articulation rate are not appropriately synthesized, so that it sounds overarticulated and with insufficient tension. It remains unclear why it concerns only the male voice.

If we just look at the three top results, then "excited" is best modelled for our female voice, whereas "competence" and "ruggedness" is best modelled for our male voice. However, it is unclear whether "excited" is associated with female and female products, and therefore supposed to be presented by a female voice. Conversely, it can be speculated whether "competent" is attributed by consumers to "male products", and consequently they are supposed to be presented by a male voice.

As a side effect, the results of this research also helped to improve the general acceptability of the diphone synthesized speech used here. We can assume that a default synthesized voice should sound sincere, competent and sophisticated, but not necessarily excited and rugged. Our results revealed that all models except for "rugged" nearly always scored better

than the baseline version for the male voice. Similarly, for the female voice the "sincere"-version was nearly always better than the baseline version. The common features for all these versions, both the male and the female voice, were an enlarged pitch range and an articulation rate that is faster than the default setting.

3. Conclusions and Outlook

We have shown that modelling personality in synthetic speech is possible. This was also demonstrated for one human personality dimension [10] and has been extended to five brand personality dimensions in this study. Nevertheless we are still only at the beginning. More research is needed, e.g. with regard to "excitement" which is also an important factor in modelling emotional synthesis, and which plays an important role in advertising and marketing.

In order to model personality traits for speech synthesis the advantage of diphone synthesis, in contrast to database synthesis, lies in the easy parameterization of the prosodic properties of the speech signal. The findings of investigations with parameterized speech give us an indication of prosodic voice parameters required for conveying a given personality impression. This can be taken into account when creating databases for unit selection if the goal is to provide one or several well-defined personalities with the same voice.

We showed that parametrical synthesis can be useful as a tool for basic research on personality aspects in speech. But it can also be used for applications such as anthropomorphized talking objects [24], speech prostheses for voice-handicapped persons, or tuning a synthetic corporate voice which suits the personality of the organisation or brand better.

4. References

[1] Aaker, J.L., 1997. Dimensions of brand personality. *Journal of Marketing Research* 34, 347-356.

[2] Apple, W., Krauss, R.M., 1979. Effects of pitch and speech rate on personal attributions. *Journal of Applied Social Psychology* 37, 715-727.

[3] Brown, B.; Strong, W.; Rencher, A., 1975. Acoustic determinants of the perceptions of personality from speech. *International Journal of the Sociology of Language* 6, 11-32.

[4] Cahn, J.E., 1990. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society* 8, 1–19.

[5] Dutoit, T.; Pagel, V.; Bataille, F.; van den Vreken, O., 1996. The MBROLA project: towards a set of quality speech synthesizers free of use for non-commercial purposes. Proc. *International Conference on Spoken Language Processing*, Philadelphia, 1393-1396.

[6] Eide, E.; Aaron, A.; Bakis, R.; Hamza; W.; Picheny, M.; Pitrelli, J., 2004. A corpus-based approach to <ahem/> expressive speech synthesis. *Proc. 5th ISCA Speech Synthesis Workshop.*

[7] Kehrein, R., 2002. The prosody of authentic emotions. Proc. *Speech Prosody 2002*. Aix-en-Provence, France.

[8] Laukka, P.; Juslin, P.N.; Bresin, R., 2005. A dimensional approach to vocal expression of emotion. *Cognition and Emotion* 19(5), 633-653.

[9] McRae, R.R.; John, O.P., 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60, 175-215.

[10] Nass, C.; Lee, K.M., 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7(3), 171-181.

[11] North, A.C.; MacKenzie, L.C.; Hargreaves, D.J., 2004. The effects of musical and voice "fit" on responses to advertisements. *Journal of Applied Social Psychology* 34, 1675-1708.

[12] Ofuka, E.; McKeown, J.D.; Waterman, M.G.; Roach, P.J. 2000. Prosodic cues for rated politeness in Japanese speech. *Speech Communication* 32, 199-217.

[13] Ohala, J.J., 1994. The frequency codes underlies the sound symbolic use of voice pitch. In: Hinton, Nichols & Ohala (eds), *Sound symbolism.* Cambridge: Cambridge University Press. 325-347.

[14] Reeves, B.; Nass, C., 1996. *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press.

[15] Robinson, J.; McArthur, L., 1982. Impact of salient vocal qualities on causal attribution for a speaker's behavior. *Journal of Personality and Social Psychology* 43, 236-247.

[16] Scherer, K. R., 1978. Personality inference from voice quality: the loud voice of extroversion. *European Journal of Social Psychology* 8, 467–487.

[17] Schmidt, S., 2005. *Persönlichkeitsaspekte in Stimmen von anthropomorphen Objekten*. Diploma thesis, Computer Science, Saarland University.

[18] Schröder, M., to appear. Expressing degree of activation in synthetic speech. to appear in: *IEEE Transactions on Speech and Audio Processing.*

[19] Schröder, M.; Grice, M., 2003. Expressing vocal effort in concatenative synthesis. *Proc. 15th International Conference of Phonetic Sciences*, Barcelona, 2589-2592.

[20] Schröder, M.; Trouvain, J., 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology* 6, 365-377.

[21] Smith, B.; Brown, B.; Strong, W.; Rencher, A., 1975. Effects of speech rate on personality perception. *Language and Speech* 18, 145-152.

[22] Streeter, L. ; Krauss, R.; Geller, V. ; Olson, C. ; Apple, W., 1977. Pitch changes during attempted deception. *Journal of Personality and Social Psychology* 35, 345-350.

[23] Trouvain, J.; Barry, W.J., 2000. The prosody of excitement in horse race commentaries. Proc. *ISCA-Workshop on "Speech and Emotion"*, Newcastle, Northern Ireland, 86-91.

[24] Wasinger, R. ; Wahlster, W., 2005. The anthropomorphized product shelf: symmetric multimodal human-environment interaction. In: Aarts, E.; Encarnaçao, J. (eds): *True Visions: Tales on the Realization of Ambient Intelligence*, Springer.

[25] Zuckerman, M.; Kunitate, M., 1993. The attractive voice: What makes it so? *Journal of Nonverbal Behavior* 17, 119 – 135.