Towards perceptual dimensions of speakers' voices: Eliciting individual descriptions

Benjamin Weiss[1]*, Felix Burkhardt*[2]*, Matthias Geier*[3]

[1]Quality & Usability Labs, Telekom Innovation Laboratories, TU Berlin, Germany
[2]Telekom Innovation Laboratories, DTAG, Berlin, Germany
[3]Institut für Nachrichtentechnik, Universität Rostock, Germany

Benjamin.Weiss@tu-berlin.de, Felix.Burkhardt@telekom.de, Matthias.Geier@uni-rostock.de

Abstract

For the small EmoDB subset of 10 neutrally recited sentences by 10 speakers, individual listeners' descriptions of voice and speech style were elicited by the method of repertory grid. With this method, one dissimilar stimulus is chosen from triples in order to create a dissimilarity space, but also to construct opposite pairs for ratings scales, individual to each of 20 listeners. In a second step, each listener rated all 10 speakers on his/her rating scales. The data obtained was compared to results from a fixed questionnaire available for the same data-base. The results show a concordance between external likeability ratings and the first dimension of perceptual distances between speakers obtained by multidimensional scaling of the dissimilarity frequencies, but only for male speakers concerning the valence of the rating scales elicited. For female speakers, the highest correlation of the first dimension is with the questionnaire factor "relaxed". Additionally, the existing questionnaire could be extended to a revised version by including attributes elicited by multiple listeners.

Index Terms: voice, timbre, perceptual dimensions

1. Introduction

In conversation, interlocutors permanently assess and rate each other on pronunciation, way of speaking, and voice, regarding paralinguistic and extralinguistic information. Several features of the interlocutor are attributed to the speaker based on such auditory information. This process is particularly important when people have to rely on little information, because the interlocutor is unknown or there is no visual channel. Such characterizations include physical traits (like gender, age, height), social traits (status or heritage), and psychological traits (e.g., emotions and personality) (cf. [1]). Traits of interpersonal relationship have to be considered as well (e.g., interest, likeability or benevolence) [2, 3]

Some of these traits can be inferred from auditory information with high validity and reliability (e.g., gender or age [4, 5]), whereas others seem to be dominated by stereotypes, providing only high reliability, and there are of course traits that cannot be inferred well (e.g., weight or height [6, 7]). Current approaches of machine learning even provide classifiers to automatically infer such information [8].

Still, many of the traits mentioned are complex constructs and therefore difficult to relate to acoustic parameters directly. One approach to deal with this issue is to attempt a preference mapping by describing such traits first with perceptual low-level aspects of voice and speaking style and to relate those well-founded attributes to the acoustic level later. However, a comprehensive view on how non-experts perceive and describe speakers in contrast to experts' schemes [9, 10] is still missing.

In [1] there are three approaches described for the purpose of solving this "dilemma" (p. 5), when assessing listeners' judgments:

Expert sets Questionnaires are built up based on list of attributes describing voices from a perceptual point of view. These lists are often based on introspection and not on theory and also the attributes of one list mostly belong to different descriptive classes (e.g., "high", "sweet", or "rigid"). The subsequent factor analyses usually include factors of the typical four dimensions of speech prosody, i.e. pitch, tempo, intensity, and timbre. Additionally, precision in pronunciation and variability in stress location and pitch is often found. As stated in [1] however, the results of course depend highly on the items of the questionnaire used. If considering more abstract attributes, for example, naturalness of speaking or aspects of personality and background, quite different factors will result from the analysis. The strength of the method are the well-defined labels for the scales and the relatively economic way to obtain comparable results.

Multidimensional scaling With this technique, stimuli are rated according to their (dis)similarity without relying on any specific attribute. This represents a major benefit, as non-expert listeners do not have to elaborate on their personal and often subconscious references when perceiving and rating stimuli. However, this complicates the interpretation of the results afterwards. By subsequent multidimensional scaling (MDS), also a dimension reduction of the data can be obtained.

Speech synthesis The third method proposed refers to the usage of voice synthesis to let participants adjust a stimulus until it perceptually fits a target stimulus best. The benefit lies in the direct comparison and instrumentally defined degrees of freedom: User variability should be reduced and thus the resulting dimensions should be more reliable. However, preparing and validating such a method (e.g., the parameters to be varied) is as complex as validating a questionnaire and will be limited to small changes to avoid artifacts during re-synthesis. Therefore, this method is not considered here.

The aim of this paper is to find relevant perceptual dimensions for unknown speakers and fitting verbal descriptions of them. The Repertory Grid Technique (RGT) is used for this purpose to complement results from an existing questionnaire (expert set) that uses a closed set of attributes [11]. RGT method

(presented in Section 3) combines benefits of multidimensional scaling and expert sets: On the one hand, individual attributes are elicited to validate and complement the scales of the questionnaire by frequent personal items. On the other hand, the ranking of voices according to the ratings obtained by RGT are used to check, which factor of the questionnaire is the most important one to distinguish the speakers. The available factors are found for the questionnaire from [11] and listed in Section 2.

2. Related Work – Material

The EmoDB contains recordings of 5 female and 5 male speakers colloquially reciting 10 German sentences with neutral content [12] (e.g. "*She wants to submit it on Wednesday.*") in various acted emotional states. These sentences do not include hesitations or repairs. From those recordings only stimuli spoken in a emotionally neutral way were chosen for this experiment.

In the aforementioned previous study [11], a questionnaire was used to let 46 listeners rate the same speech material produced emotionally neutral as in the paper at hand. It was based on items from [13, 14, 15] and validated in two small focus group sessions with five experts each (phoneticians and psychologists). The questionnaire is divided into two part: A smaller part contains scales to assess an *impression of the speaker* (13 items, e.g. likeability, self-assurance), as well as a longer part related to the *voice and speaking style* (25 items, e.g. resonance, roughness). A preliminary analysis of the likeability data only, used for classifying, can be found in [16].

A factor analysis resulted in five factors for the *speaker traits* Likeability, Activity, Dominance, Attractiveness and Reservation (only one item) and seven factors for the *voice and speaking style* (Relaxed, Dark, Strong, Inconspicuous, Professional, Fluent and Voluminous (only one item)) [11].

Due to the limitations of expert sets, this study uses the complementing method of RGT described below in order to validate and enhance the questionnaire used in [11].

3. Repertory Grid Technique – Method

RGT is a method from the area of personal construct psychology [17]. It has been used in the field of acoustics, e.g., for perceptual analysis of textural sounds [18] or spatial sound perception [19, 20], but also for the analysis of stuttering [21].

The procedure is divided into two parts, an elicitation phase and a rating phase, in our case implemented as computer program operated by mouse and keyboard. Basically, participants are asked in the first phase to find similar and dissimilar elements and name the characteristics of (dis-)similarity. Ideally, each participant can thus build her own antonyms (called constructs) used in the second phase to rate all elements. The main idea of RTG is to make explicit individual constructs or aspects associated with the elements studied. We are not interested in the idiosyncrasy of attributions of voices and speaking style, but only in the reoccurring concepts and labels of speakers. RTG avoids the issue of pre-defined labels of expert sets and allows for MDS analysis to find major perceptual dimensions, i.e. those used for deciding on overall similarity.

In this particular case, the so-called "elements" have been pre-defined as the 10 speakers, but the "constructs" (i.e. attributes) to describe and distinguish the speakers are individually elicited by each listener. This combination is called "partial RGT" according to [22]. Various sentences (stimuli) are chosen to represent each speaker (i.e. element) in order to avoid boredom of the listeners by varying the stimuli, although sentence content was controlled within each set of presented stimuli.

1. During the *elicitation phase*, triples of stimuli were chosen to be presented to each listener, who was asked to identify the pair of voices (stimuli) to be more similar contrasting the dissimilar stimulus. Furthermore, the listener had to describe as free text in what way the pair is similar (e.g., both "fast"), and how it is different from the third stimulus (e.g., "slow"), *thus eliciting a pair of (assumed) opposite attributes* for each triple. If necessary, a second pair could be provided. The participants should also indicate whether the first attribute of the pair is considered negative, positive, or neutral. This is not typically part of RGT, but successfully used in [20]. Gender was separated, as a first block of triples was presented exclusively from either male or female speakers, followed by a block of the opposite gender. Within each triad, the sentence content was similar for each stimulus, but varied between triads. Altogether, 10 triads were presented for each gender in order to reach a complete set of possible speaker triples, resulting in 60 different stimuli of the 100 available from the EmoDB. These 60 stimuli have been randomized in order to create individual playlists for each participants covering all 100 stimuli in a roughly balanced way.
2. After a short break, during which misspellings and repetition have been removed manually from the list of attribute pairs, the listener had to rate all speakers according to each of his/her attribute pairs provided in the *rating phase*. A slider was used for this rating. The stimuli were not presented as triples, but all five speakers of one gender uttering the same sentence presented on one screen. Here, gender was alternated, so that a listener could rate all 10 speakers on one scale on two consecutive screens. The sentence content was randomized for each screen, i.e. between genders and attribute pair.

AKG K-601 headphones were used for playback. 20 listeners without expert background regarding psychology, phonetics, or the like participated in this experiment (aged 21–37, $M = 25.9$, $SD = 5.2$, gender balanced). They have been paid for their contribution. One complete session took about one hour.

4. Results

The results for this RGT experiment are presented in three parts, the elicited individual attribute pairs, the similarity ratings (both from the first phase), and the actual numerical ratings of the stimuli on each individual scale (from the second phase).

4.1. Attribute Pairs

After removing duplicates, each of the 20 listeners used 3–21 attribute pairs ($M = 14.5$, $SD = 5.4$). Not all of the pairs resemble true opposites, but could be interpreted nevertheless (e.g. high pitch–low pitch, similar accentuation–different accentuation, but also: interested–suspicious; understanding–brief; thoughtful–cold).

In order to collect the main frequent aspects used for labeling the speaker (dis-)similarities, factor analysis was conducted on each individual grid. After naming each factor adopting or rephrasing individual attributes, the following ten attribute pairs represent factors occurring more than once: *similar sentence stress–different sentence stress; fast–slow; high*

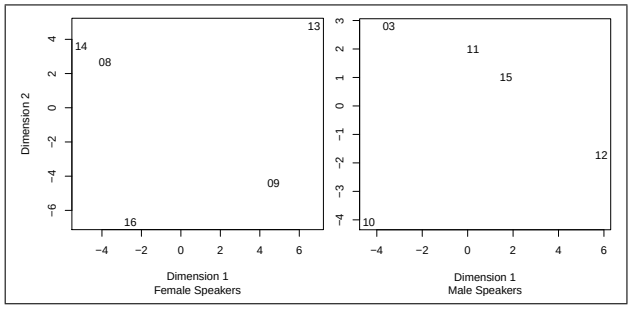

Figure 1: *2-dimensional MDS solution for the dissimilarity frequencies of the elicitation phase. Numbers identify speakers from the database.*

pitch–low pitch; calm–excited; factual–emotional; interested–reserved; precise pronunciation–imprecise pronunciation; pleasant timbre–unpleasant timbre; positive emotion/friendly–negative emotion/unfriendly; loud–gentle.

Interestingly, the last pair occurred only twice, which could be a result of the limited variability due to the recording instructions for the neutral session. Concerning the attribute *emotional/friendly*, there have been various positive and negative descriptions used. But the highest variability was with the attributes related to voice or timbre (e.g., *pleasant timbre*: soft/warm/neutral/full voice vs. *unpleasant timbre*: hard/cold/shrill/small voice).

Attributes currently not included in the questionnaire are those related to sentence stress location and factual style. The questionnaire currently offers *interesting–boring* to describe a speaker. In the elicitation, however, *interested* was used (contrasted to *uninterested, bored*). Therefore, this attribute pair should be revised to cover interest expressed by voice instead of an intriguing, interesting speaker.

4.2. (Dis)Similarity Ratings

The data from the elicitation phase was aggregated over the sentences to extract the differences between the speakers. The Euclidean distance between speakers of each gender was calculated by using the frequencies of dissimilar pairs. Applying MDS on these distances, a clear separation of the speakers is found, stronger for the female than for the male group (cf. Figure 1).

On the basis of the 2-dimensional solution, the distances are compared to the factor values of the speakers from the fixed questionnaire [11]. Based on the average values of the speakers correlations are strongest between Dimension 1 and *Speaker: Likeability (-.74) and Activity (.72)* and MDS Dimension 2 and *Voice: Professional* for the female speakers, as they are between MDS Dimension 1 and *Speaker: Likeability* and MDS Dimension 2 and *Speaker: Attractiveness* for males (cf. Figure 2). See Table 1 for the coefficients.

The differences in sign for MDS 1 with female speakers and Likeability and Activity may indeed indicate that higher activity is perceived negatively. Although there is no significant correlation between these two dimensions, a relaxed voice is strongly correlated with a likeable speaker (cf. [11]).

4.3. Stimulus Ratings

From the actual ratings of all speakers during the rating phase, data from non-neutral attribute pairs was taken, as indicated by the participants during the elicitation phase. This data was poled to average the ratings and subsequently rank each speaker.

From altogether 578 attribute pairs, 394 (68%) were marked as either positive or negative. The ratings on scales with valence were z-transformed for each listener and then averaged for each speaker and listener to obtain a comparable weighting for the participants. Then, the mean over all listeners was calculated for each speaker. See Table 2 for the resulting ranking of speakers.

Table 1: *Pearson's correlations ($r \geq .7$) between both MDS dimensions and factors from [11].*

	women		men	
	MDS 1	MDS 2	MDS 1	MDS 2
Speaker: Likeability	.96		-.74	
Speaker: Activity			.72	
Speaker: Dominance	.82			
Speaker: Attractiveness				-.73
Voice: Professional		.92		

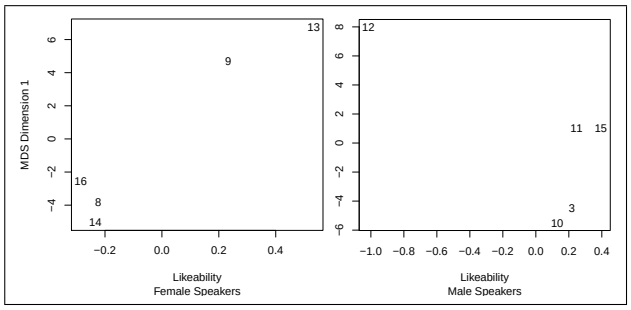

Figure 2: *Correlation between speaker distances and the Likeability factor of the questionnaire results from [11]. Numbers identify speakers from the database.*

Table 2: *Ranking of the speaker according to the standardized non-neutral ratings.*

women	value	men	value
13	-0.172	15	-0.120
14	-0.119	11	-0.085
09	-0.045	03	-0.032
08	0.017	10	0.034
16	0.244	12	0.276

Using correlation analysis as in the last section, these valence values show concordance with data from [11], namely the highest correlation with *Speaker: Likeability* ratings for male speakers and with the factor *Voice: Relaxed* for female speakers (cf. Figure 3 and Table 3).

5. Discussion

When comparing the rating results for the same material obtained by the RGT method to those obtained with a fixed questionnaire, similarities can be found for the underlying dimensions of the RGT and the factors of the questionnaire. The

Table 3: *Pearson's correlations ($r \geq .7$) between ratings of attributes with valence and factors from [11].*

	women	men
Speaker: Likeability		-.97
Speaker: Activity		
Speaker: Dominance		
Speaker: Attractiveness		
Voice: Relaxed	-.88	-.84
Voice: Dynamic	-.82	-.75
Voice: Professional	-.81	-.78
Voice: Fluent	-.75	-.93

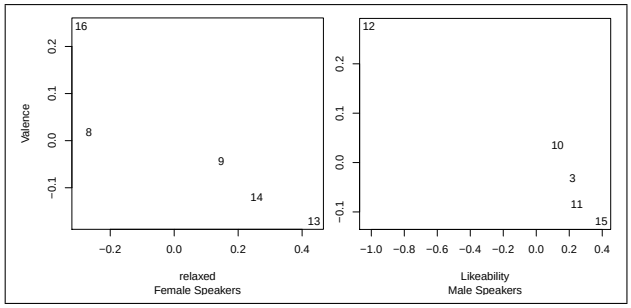

Figure 3: *Correlation between Valence and Factors of the questionnaire results from [11]. Numbers identify speakers from the database.*

primary dimension of calculated distances between the speakers, which are based on dissimilarity ratings, relate to distances between *Likeability* ratings of these persons for both male and female speakers. The second dimension is different, however: *Professionalism of speaking* for women and *Attractiveness* for men is most strongly correlated to the secondary dimension. For the few speakers compared, overall preference described as likeability seems to be the most relevant factor to differentiate them.

A similar analysis using the ratings on scales with non-neutral personal constructs confirms the relevance of likeability to distinguish unknown speakers only for the five men. The valence ratings of the female speakers correlate strongest with the questionnaire factor *Relaxed voice*.

The questionnaire factor Likeability subsumes questionnaire attributes "friendly", "likeable", "sympathetic" and "pleasant" and might represent the first of the two major dimensions of person perception "benevolence" and "competence" (cf. [23, 24]).

As the number of speakers tested is very low, no generalizations to other speakers can be drawn from the results obtained. But still, the concordance with likeability is very plausible and shows that, for the participants and material presented here, likeability constitutes a major perceptive or evaluative dimension to differentiate and assess speakers according to their voices. Other perceptual factors, e.g. *Relaxed*, do also correlate with these distances and ratings. However, for the task given to the participants, it seems that those perceptual concepts closer related to voice and speaking style might be used more indirectly in favor of the higher level interpersonal rating of *likeability* to distinguish unknown speakers.

With more data, attempting a preference mapping of those more descriptive concepts of voice and speaking to interpersonal ratings, as stated in the introduction, seems even more promising.

The RGT method itself has to be considered very laborious, at least in this form of complete sets of triples concerning the speaker pairings during the elicitation phase. An adapted version suitable for more speaker variation is needed to obtain distances that can be generalized over speakers.

In order to elicit promising questionnaire items for a representative selection of speakers, it seems advisable to use only the first part, the elicitation phase, for randomized stimulus triples, as data from the rating phase are very difficult to analyze for different stimulus sets between the participants. For example, we did not present any of the 20 individual grids (the results of the ratings phase), as 20 separate factor or principal component analyses would not be meaningful.

Although most of the frequent attributes are somehow already included in the questionnaire, the elicitation phase still gave new ideas on how non-expert listeners describe unknown voices and the questionnaire presented in [11] will be enhanced as stated in Section 4.1: The frequently mentioned attributes *factual–emotional* and *typically stressed–peculiarly stressed* will be added. Also, *interested–not interested* will be used instead of *interesting–boring* for further validation. Still, this result does not indicate a roughly "complete" set of attributes, but this methods is one way to validate the questionnaire concerning one major issue of limited experts sets.

6. Conclusions

The repertory grid technique was used as one step to identify perceptual dimensions of speakers' voices. Distances between five male and five female speakers are strongest related to likeability ratings, obtained as a resulting factor from a questionnaire. The individual attributes to describe differences in speakers based on their voices and speaking style can be used to supplement other means in the creation and validation of suitable questionnaires assessing features of speakers.

7. Acknowledgements

This work was funded by the German Research Foundation DFG (WE 5050/1-1). We would also like to thank Charlotte Buchsbaum for her help with the experiment and the anonymous Reviewers for their valuable comments.

8. References

[1] J. Kreiman and D. Sidtis, *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Chichester: Wiley, 2011.

[2] A. Mehrabian, "Some referents and measures of nonverbal behavior," *Behavioral Research Methods and Instrumentation*, vol. 1, pp. 213–217, 1969.

[3] B. L. Smith, B. L. Brown, W. J. Strong, and A. C. Rencher, "Effects of speech rate on personality perception," *Language and Speech*, vol. 18, pp. 145–152, 1975.

[4] F. Burkhardt, R. Huber, and A. Batliner, "Application of speaker classification in human machine dialog systems," in *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller, Ed. Springer, 2007, pp. 174–179.

[5] R. Winkler, "Merkmale junger und alter Stimmen – Analyse ausgewählter Parameter im Kontext von Wahrnehmung und Klassifikation," Dissertation, Technische Universität Berlin, Berlin, 2009.

[6] W. van Dommelen and B. Moxness, "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," *Language and Speech*, vol. 38, pp. 267–287, 1995.

[7] J. Cohen, T. Crystal, A. House, and E. Neuberg, "Weighty voices and shaky evidence: A critique," *Journal of the Acoustical Society of America*, vol. 68, pp. 1884–1885, 1980.

[8] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2013.

[9] J. Laver, *The phonetic description of voice quality*. Cambridge: University Press, 1980.

[10] M. Hirano, *Clinical Examination of Voice*. New York: Springer, 1981.

[11] B. Weiss and S. Möller, "Wahrnehmungsdimensionen von Stimme und Sprechweise," in *22th Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Aachen*, ser. Studientexte zur Sprachkommunikation, B. Kröger and P. Birkholz, Eds., vol. 61. Dresden: TUDpress, 2011, pp. 261–268.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[13] K. Scherer, "Personality inference from voice quality: The loud voice of extraversion," *European Journal of Social Psychology*, vol. 8, pp. 467–487, 1978.

[14] B. Ketzmerick, *Zur auditiven und apparativen Charakterisierung von Stimmen*, ser. Studientexte zur Sprachkommunikation. Dresden: TUDpress, 2007.

[15] S. Singh and T. Murry, "Multidimensional classification of normal voice qualities," *Journal of the Acoustical Society of America*, vol. 64, no. 1, pp. 81–87, 1978.

[16] B. Weiss and F. Burkhardt, "Voice attributes affecting likability perception," in *Proc. INTERSPEECH*, 2010, pp. 1934–1937.

[17] G. Kelly, *The Psychology of Personal Constructs*. New York: Norton, 1955.

[18] T. Grill, A. Flexer, and S. Cunningham, "Identification of perceptual qualities in textural sounds," in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, 2011.

[19] J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Proceedings of the 16th AES Conference*, 1999.

[20] M. Geier, H. Wierstorf, J. Ahrens, I. Wechsung, A. Raake, and S. Spors, "Perceptual evaluation of focused sources in wave field synthesis," in *Audio Engineering Society Convention, London*, 2010.

[21] F. Fransella, R. Bell, and D. Bannister, *A Manual for Repertory Grid Technique*, 2nd ed. Chichester: Wiley, 2004.

[22] H. M. Edwards, S. McDonald, and S. M. Young, "The repertory grid technique: Its place in empirical software engineering research," *Information and software technology*, vol. 51, pp. 785–798, 2008.

[23] A. E. Abele, A. J. C. Cuddy, C. M. Judd, and V. Y. Yzerbyt, "Fundamental dimensions of social judgment. editorial to the special issue," *European Journal of Social Psychology*, vol. 38, no. 7, pp. 1063–1065, 2008.

[24] M. Schaller, "Evolutionary basis of first impressions," in *First Impressions*, N. Ambady and J. J. Skowronski, Eds. New York: Guilford Press, 2008, pp. 15–34.